Ph.D. DISSERTATION

Image Registration by Feature Blending Network with Applications in High Dynamic Range Imaging and Video Super-resolution

특징 혼합 네트워크를 이용한 영상 정합 기법과 고명암비 영상법 및 비디오 고 해상화에서의 응용

BY

LEE SANG-HOON

AUGUST 2020

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Ph.D. DISSERTATION

Image Registration by Feature Blending Network with Applications in High Dynamic Range Imaging and Video Super-resolution

특징 혼합 네트워크를 이용한 영상 정합 기법과 고명암비 영상법 및 비디오 고 해상화에서의 응용

BY

LEE SANG-HOON

AUGUST 2020

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Image Registration by Feature Blending Network with Applications in High Dynamic Range Imaging and Video Super-resolution

특징 혼합 네트워크를 이용한 영상 정합 기법과 고명암비 영상법 및 비디오 고 해상화에서의 응용

지도교수 조 남 익

이 논문을 공학박사 학위논문으로 제출함

2020년 8월

서울대학교 대학원

전기 컴퓨터 공학부

이 상 훈

이상훈의 공학박사 학위 논문을 인준함

2020년 8월

위 원 장: _____
부위원장: _____
위　　원: _____
위　　원: _____
위　　원: _____

Abstract

This dissertation presents a deep end-to-end network for high dynamic range (HDR) imaging of dynamic scenes with background and foreground motions. Generating an HDR image from a sequence of multi-exposure images is a challenging process when the images have misalignments by being taken in a dynamic situation. Hence, recent methods first align the multi-exposure images to the reference by using patch matching, optical flow, homography transformation, or attention module before the merging. In this dissertation, a deep network that synthesizes the aligned images as a result of blending the information from multi-exposure images is proposed, because explicitly aligning photos with different exposures is inherently a difficult problem. Specifically, the proposed network generates under/over-exposure images that are structurally aligned to the reference, by blending all the information from the dynamic multi-exposure images. The primary idea is that blending two images in the deep-feature-domain is effective for synthesizing multi-exposure images that are structurally aligned to the reference, resulting in better-aligned images than the pixel-domain blending or geometric transformation methods. Specifically, the proposed alignment network consists of a two-way encoder for extracting features from two images separately, several convolution layers for blending deep features, and a decoder for constructing the aligned images. The proposed network is shown to generate the aligned images with a wide range of exposure differences very well and thus can be effectively used for the HDR imaging of dynamic scenes. Moreover, by adding a simple merging network after the alignment network and training the overall system end-to-end, a performance gain compared to the recent state-of-the-art methods is obtained.

This dissertation also presents a deep end-to-end network for video super-resolution (VSR) of frames with motions. To reconstruct an HR frame from a sequence of adjacent frames is a challenging process when the images have misalignments. Hence,

recent methods first align the adjacent frames to the reference by using optical flow or adding spatial transformer network (STN). In this dissertation, a deep network that synthesizes the aligned frames as a result of blending the information from adjacent frames is proposed, because explicitly aligning frames is inherently a difficult problem. Specifically, the proposed network generates adjacent frames that are structurally aligned to the reference, by blending all the information from the neighbor frames. The primary idea is that blending two images in the deep-feature-domain is effective for synthesizing frames that are structurally aligned to the reference, resulting in better-aligned images than the pixel-domain blending or geometric transformation methods. Specifically, the proposed alignment network consists of a two-way encoder for extracting features from two images separately, several convolution layers for blending deep features, and a decoder for constructing the aligned images. The proposed network is shown to generate the aligned frames very well and thus can be effectively used for the VSR. Moreover, by adding a simple reconstruction network after the alignment network and training the overall system end-to-end, A performance gain compared to the recent state-of-the-art methods is obtained.

In addition to each HDR imaging and VSR network, this dissertation presents a deep end-to-end network for joint HDR-SR of dynamic scenes with background and foreground motions. The proposed HDR imaging and VSR networks enhace the dynamic range and the resolution of images, respectively. However, they can be enhanced simultaneously by a single network. In this dissertation, the network which has same structure of the proposed VSR network is proposed. The network is shown to reconstruct the final results which have higher dynamic range and resolution. It is compared with several methods designed with existing HDR imaging and VSR networks, and shows both qualitatively and quantitatively better results.

keywords: Image registration, Convolutional neural network, High dynamic range imaging, Video-super resolution

student number: 2013-20846

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Dynamic ranges of standard cameras are too narrow when compared with those of most scenes around us. Also, they cannot capture too bright or dark regions that have illumination values out of the ranges of normal camera-settings. Thus, high dynamic range (HDR) imaging, which is a technique to capture and express HDR scenes in a single image, has been studied to overcome the limitation. The most common approach is to take a sequence of low dynamic range (LDR) images with different exposures and fuse them to an HDR image [9,18]. Then, the HDR image is appropriately tonemapped to the display dynamic range [11, 53, 64]. Another approach is directly synthesizing a tonemapped-like image as a weighted sum of LDR images, which is called exposure fusion [46, 52, 57].

However, the classical HDR imaging methods produce plausible results only when the camera is fixed on a tripod, and when all the objects in the scene do not move, which are too impractical or limited conditions. In a practical situation, there are misalignments in the multi-exposure photos, due to background and foreground motions. If the misaligned photos are merged into an HDR image, most regions in the result are blurred due to the background motion, and there can be ghost artifacts around the moving objects. To alleviate these problems, advanced HDR imaging methods include the algorithm to align the input images to the reference that is usually the image with

medium exposure. For some examples, the alignment is implicitly or explicitly conducted using patch matching [25], optical flow [33], homography transformation [78], or attention module [79]. However, the alignment methods do not produce satisfying results when there are complicated background differences, nonrigid motion, or illumination difference, which results in low-quality results, sometimes with fatal artifacts. Thus, it is still needed to improve the alignment step before merging the multi-exposure images because the alignment quality is critical for reducing the blurring and ghost artifacts. In Fig. 1.1, there are two images with different exposures, the source and reference, and they have nonrigid motion. The aligned result using the existing alignment method using optical flow has fatal artifacts, while the result by the proposed method doesn't.

Super-resolution (SR) is to enhance resolution of an low-resolution (LR) image by reconstructing fine details in the image. Consider a sequence of LR video frames frames, a target frame among them can be reconstructed by using single image super-resolution (SISR). However, this is wasteful method, since details available from the neighbor frames are not used. The frames include different details in the same regions, and they are very useful to enhace the resolution. In video super-resolution (VSR) or multi image super-resolution (MISR), the details included in the neighbor frames are fused for super-resolution.

To extract the details in the neighbor frames isn't dificult task when there are no motions between the frames, and they are spatially aligned perfectly, which is a too slim situation. In practical situation, however, the frames in the sequence have both background and foreground motions and the frames should be aligned explicitly or implicitly to extract the details. To align the consecutive frames, existing VSR imaging methods include the method to align the neigbor frames to the target frame. For example, the alignments are usually conducted using spatial transformer networks (STN) which are based on optical flow estimation [6,21,47,67]. However, the alignment methods do not produce satisfying results, and sometimes cause with fatal artifacts.when

(a) Source (b) Reference

(c) Aligned Result (Optical Flow) (d) Aligned Result (Proposed)

Figure 1.1: (a) The source image with low exposure. (b) The reference image with standard exposure. (c) The aligned result using optical flow based method. (d) The aligned result using the proposed method

there are complicated background differences or nonrigid motion. Thus, it is important to get well aligned frames before reconstruct the final high-resolution (HR) frame with the frames.

In this respect, this dissertation proposes an end-to-end convolutional neural network (CNN) that first generates well-aligned multi-exposure images and then merges the aligned images in Chapter. 3. It has been noted that explicitly aligning differently exposed images is not an easy task, as evidenced in the previous works. Thus, this dis-

sertation focus on the alignment part while a simple network is used for the merging part. For synthesizing well-aligned differently exposed images, the proposed main idea is to blend the images at the deep-feature-domain, rather than to blend or transform the images at the pixel- or shallow-feature-domain. Then the blended features are used to construct an aligned image. Precisely, a building block called Exposure-Structure Blending Network (ESBN) is proposed, which generates an under- or over-exposure image that is structurally aligned to the reference image. The ESBN consists of two encoders that extract deep features from two differently exposed images separately, several residual blocks that blend the deep features, and a decoder that constructs the aligned image. If N LDR exposure images are given, then $N-1$ ESBNs are used to generate all the aligned exposure images. Then, the aligned images are fed to a simple merging network, which consists of a densely connected residual block.

In the experiment, the ESBN and merging network are first trained separately to show the effectiveness of ESBN over the conventional alignment methods. It is shown that the ESBN works well for quite large exposure differences, and thus any existing (even non-dynamic) HDR imaging methods can also use the outputs of ESBN for generating a plausible HDR image. Also, the overall network is trained end-to-end to validate a performance gain of joint alignment and merging networks. From the extensive qualitative and quantitative comparisons, it is shown that separate training of ESBN and merging network produces better results than existing state-of-the-art methods, and the end-to-end training yields even better performance.

In Chapter. 4, an end-to-end CNN that first aligns neighbor frames and then reconstructs the HR frame by merging the aligned frames is proposed. It has been noted that generating aligned frames elaborately is not an easy task, as evidenced in the previous works. To align the neighbor frames well, this dissertation proposes the method in common with the method in Chapter. 3, which blends the images at the deep-feature-domain, rather than to blend or transform the images at the pixel- or shallow-feature-domain. By modifying the ESBN, another building block called

Detail-Structure Blending Network (DSBN) is proposed, which structurally aligns the neighbor frames to the reference target frame. The DSBN also consists of two encoders that extract deep features from two input frames separately, several residual blocks that blend the deep features, and a decoder that reconstructs the aligned result. If N consecutive frames are given, then $N-1$ neighbor frames are aligned to the target frame which is temporally centered. Then, the aligned frames are fed to a simple HR reconstruction network, which consists of several residual blocks and a decoder.

In the experiment, it is shown that the DSBN works well for any frames even with non-rigid motions, and high-quality HR frames are reconstructed with the aligned frames using the HR reconstruction network. Also, the overall VSR network which inlcudes the DSBNs and the reconstruction network is trained end-to-end to validate a performance gain. From the extensive qualitative and quantitative comparisons, it is shown that the proposed method has better performance than existing state-of-the-art methods even with lower execution time.

In Chapter. 5, this dissertation presents a method for joint HDR-SR of dynamic scenes which have different exposures and both background and foreground motions. The proposed networks in Chapter. 3 and Chapter. 4 enhance the dynamic range and the resolution of images, respectively, however they can be enhanced simultaneously by a single network. To enhace both the dynamic range and the resolution, the proposed VSR network in In Chapter. 4 is used at it is, while it is trained to reconstruct the result image with both higher resolution and dynamic range.

The proposed joint HDR-SR network is compared with various networks in the experiment. The networks are variations of the network used in the existing HDR imaging and VSR methods. Some are combinations of the existing HDR imaging and SISR networks, and others are the existing VSR networks which are retrained to generate the result image with higher dynamic range as well as higher resolution. it is shown that the proposed joint HDR-SR has better performance than other networks by evaluating qualitatively and quantitatively.

The main contribution of this dissertation is the ESBN and DSBN. The ESBN blends the exposure and structure information from differently exposed and dynamically changing images, and as a result, generates multi-exposure images that are structurally aligned to the reference. The aligned images can also be used as the inputs to existing (non-)dynamic HDR imaging methods. Likewise, the DSBN blends the detail and structure information from the frames with different missing details. The other contribution is an end-to-end HDR imaging and VSR network that combines the alignment and a simple merging network, which provides state-of-the-art performance.

Chapter 2

Related Work

2.1 High Dynamic Range Imaging

The conventional HDR imaging methods generate a final HDR image H as a weighted sum of input LDR images in the HDR domain (generally in the 32-bit RGBE format), which can be expressed as

$$H = \sum_{n=1}^{N} W_n H_n, \tag{2.1}$$

where N is the number of LDR images, H_n is the n-th input image in the HDR domain, and W_n denotes the weight corresponding to H_n. In dynamic scenes, however, the methods face limitations such as blurring or ghost artifacts due to background and/or foreground motions. To overcome the limitations, some methods detect regions with motions in H_n and then reject them by reducing W_n. Also, other methods first align H_n and then compute W_n. In recent years, patch-based optimization and CNN-based learning methods have also been proposed for the better HDR generation.

2.1.1 Rejecting Regions with Motions

The first approach is to detect regions with foreground motions in input images and reject them in the merging process. The methods with this approach assume that multi-

exposure images have no background motions or they are already aligned globally, where each of the methods has a different way of measuring the motion. For some examples, Jacobs *et al.* [29] computed the local entropy of input images to detect regions with motions. Heo *et al.* [23] and An *et al.* [1, 3] computed the correlation between the images to reject the moving regions. Zhang and Cham [86] analyzed the magnitude and orientation of gradients to classify regions with and without motions. An *et al.* [2] also attempted to find the moving regions by measuring the zero-mean normalized cross-correlation. Lee *et al.* [44] and Oh *et al.* [60] used rank minimization to find outlier regions, and Yan *et al.* [80] synthesized ghost-free HDR images based on the sparse representation. However, these methods are not appropriate for the scenes with large background motions because they ignored some of the misaligned regions that might contain valuable information for improving the results. It is essential to have aligned images to use the information from different exposures as much as possible, in contrast to the methods that discard the moving regions.

2.1.2 Alignment Before Merging

The second approach is to align input images to the reference that is usually the image with medium exposure or the best-exposed image chosen by quality measures. The images having the background and/or foreground motions compared to the reference are aligned using geometric transformation and/or optical flow. For some specific examples, Ward *et al.* [77] aligned backgrounds using translation transformation, and Tomaszewska and Mantiuk [75] used homography transformation. However, since these methods cannot deal with foreground motions, Bogoni [5] additionally used optical flow to align moving objects. Also, Kang *et al.* [35] computed the optical flow using a method in [51] to estimate local motions. Zimmer *et al.* [88] estimated the optical flow by minimizing an energy function that includes gradient and smoothness terms. Hu *et al.* [24] found regions with dense correspondences between the images to align them. These methods work well in general but not in challenging cases such

as nonrigid motion. Hence the final HDR images sometimes show the artifacts. Having well-aligned multi-exposure images is the most important for the successful HDR imaging of dynamic scenes, and thus focus on the generation of aligned multi-exposure images in the deep learning framework.

2.1.3 Patch-based Reconstruction

The third approach is to reconstruct an HDR image patch by patch, unlike previous pixel-wise methods. It first finds patches with dense correspondences between the input and reference, and then reconstructs patches of final HDR image by solving an optimization problem. Specifically, Sen *et al.* [68] reconstructed an HDR image by restoring missing information in saturated regions of the reference image using other input images. Hu *et al.* [25] proposed a patch-based reconstruction of multi-exposure images, which are then used as the input to the existing merging methods. These methods generally yield high-quality HDR images with less artifact, but they require high computational complexity due to the repetition of the optimization for all patches.

2.1.4 Deep-learning-based Methods

Based on many successful CNN-based image restoration methods, there have also been several HDR imaging methods using deep networks. They take a stack of LDR images as input and produce an HDR image using a CNN, sometimes with a pre-processing step such as the optical flow or homography transform. Precisely, Kalantari and Ramamoorthi [33] aligned input images using an optical flow method in [49] and merged the aligned images into an HDR image using a CNN. Wu *et al.* [78] first aligned background motions using homography transformation and merged the aligned images using an encoder-decoder architecture. Yan *et al.* [79] used attention modules to detect useful regions and misaligned regions, and adopted dilated residual dense blocks to merge the images. These CNN-based methods not only merge input LDR images into an HDR image but also restore the information of saturated regions. Unlike the pre-

vious methods, a network that generates an aligned image is designed along with a merging network. Specifically, the alignment network generates an image with a certain exposure to have the same structure as the one with different exposure. Compared to the previous methods that perform image alignment by optical flow or homography transform, the proposed network can be trained end-to-end along with the merging network, and hence performs faster at the inference. Also, compared to the attention module that implicitly aligns the image by focusing on similar areas across the images, the proposed method generates accurately aligned images by blending the information from multi-exposure images. Hence, the alignment network can also be used as an efficient pre-processing step for other HDR imaging methods.

2.1.5 Single-Image HDRI

Unlike previously mentioned methods that use bracketed multi-exposure images as the input, there is another approach that generates an HDR image from a single LDR input. Specifically, some of the methods belonging to this approach generate virtual multi-exposure images from the input and then blend them with the appropriate weight maps [16, 19, 31, 40, 61, 62, 72, 76]. Also, there are reverse tone mapping operators (rTMOs) that map the LDR to the HDR as presented in [4, 28, 41, 55, 56]. Recently, CNN-based single-image HDR and rTMO have also been presented in [7, 12, 13, 34, 45]. Specifically, a CNN-based exposure fusion was introduced in [34]. Also, a CNN-based rTMO was also proposed in [13], where they generate virtual multi-exposure images by using a CNN, and then fuse them for generating the HDR image. In [12], they focused on the saturated regions and predicted the contents of those regions to generate an HDR image. Also, a GAN-based inverse tone mapping was proposed in [45]. There is also a method that jointly performs HDRI and SR from a single input [70]. In summary, the purpose of single-image HDR or rTMO is to generate an HDR image from an LDR input. These methods have advantages in that there are no ghost artifacts in the output because they use a single input. However, when well-taken or

well-aligned multi-exposure images are given, multi-input HDRI methods would perform better than the single-input HDRI because more information is available from the multi-exposure inputs, especially in the dark and washed-out regions.

2.2 Video Super-resolution

Super-resolution is an ill-posed inverse problem with wide approaches ranging from simple interpolation methods such as Bilinear and Bicubic to example [14, 15, 74, 82], self-similarity [26, 81] and dictionary learning [63, 83] based approaches. Including these approaches, extensive overviews are provided by Milanfar [58] and Nasrollahi and Moeslund [59].

Deep learning has shown its superior performance in various computer vision tasks including image SR. Dong *et al.* citedong2015image firstly introduced a three-layer CNN in image SR to learn a complex LR-HR mapping. Kim *et al.* [38] increased the depth of CNN to 20 layers for more contextual information usage in LR images. In [38], a skip connection was employed to overcome the difficulty of optimization when the network became deeper. Recent studies have adopted different kind of skip connections to achieve remarkable improvement in image SR.

2.2.1 Deep Single Image Super-resolution

Deep learning have achieved dramatic advantages against conventional methods in image super-resolution. SISR using deep learning is first introduced by Dong *et al.* [10] that requires a predefined upsampling operator and proposed an end-to-end mapping between the interpolated LR input and their HR output. This baseline have been further improved mainly by increasing network depth or sharing network weights. Kim *et al.* [38] and Zhang *et al.* [85] increase the depth of the network by using more convolutional layers with residual learning. Kim *et al.* [39] also introduce recursive learning in a very deep network for parameter sharing, and Tai *et al.* [71] propose recursive blocks

for very deep networks. Furthermore, improvements using up-sampling layers [69], back-projection [20], progressive up-sampling [42] and residual dense block [87] are introduced.

2.2.2 Deep Video Super-resolution

In VSR, the basic approach to refer temporal information is to concatenate the frames as in [6, 36, 48]. They are based on SISR, while they accept multiple input images. Jo *et al.* [30] propose to estimate dynamic up-sampling filters and residual images, however it fails to deal with multiple motions in the consecutive frames. To deal with the problems with multiple motions, Liu *et al.* [50] introduce multiple SR inferences for different motions. The multiple branches in their network generate their own outputs, and the final output is constructed by aggregating the outputs. Bidirectional RNN is first used by Huang *et al.* [27], and Tao *et al.* [73] add motion compensation modules. Recently, Sajjadi *et al.* [67] propose an efficient RNN based method that uses the previous HR estimation result to generate the next HR frame.

Chapter 3

High Dynamic Range Imaging

3.1 Motivation

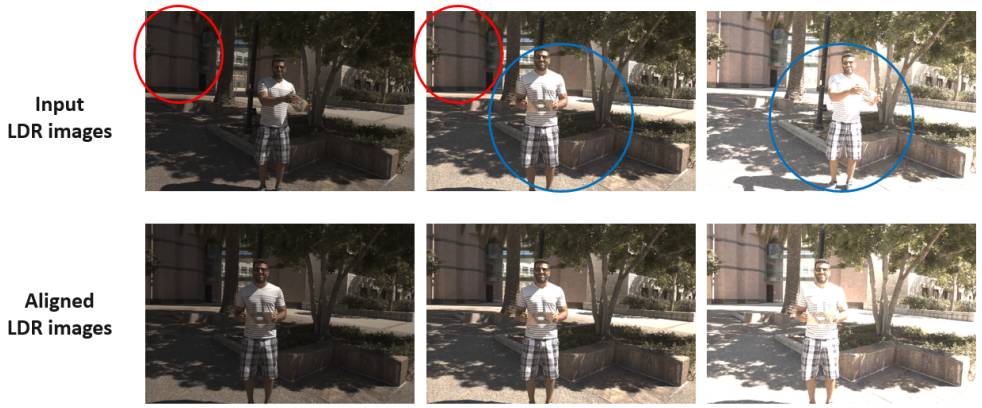


Figure 3.1: The input LDR images before which have both background (red circles) and foreground motions (blue circles), and the aligned result.

In practical situations, there are misalignments in the multi-exposure images, due to background and foreground motions as shown in Fig. 3.1. Before directly merging the images, the input images are should be aligned. Some recent HDR imaging methods explicitly align input LDR images using optical flow [33] or homography transformation [78]. Both methods align the background motions well, but the former

produces some artifacts in the nonrigid foreground objects. The latter has the limitation that it can align only one homography transformation, *i.e.,* the homography either for background or a large planar foreground object. In addition to these shortcomings, they are difficult to be parallelized and take long computation times. For example, the alignment process using the optical flow in [33] takes a half minute approximately. Thus, this dissertation proposes a fast and elaborate alignment method which synthesizes aligned images using a CNN.

3.2 Proposed Method

3.2.1 Overall Pipeline

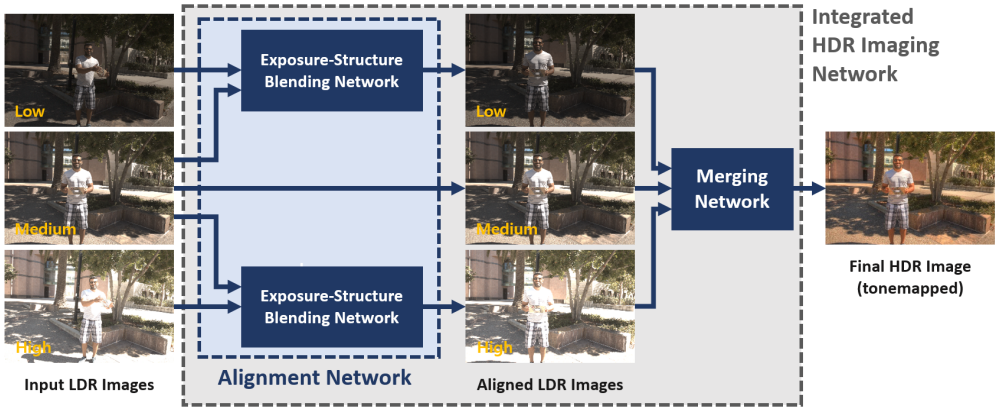


Figure 3.2: The overall pipeline of the proposed HDR imaging method. First, structurally aligned multi-exposure images are generated in the Alignment Network, which consists of exposure-structure blending networks (ESBNs). Then, the aligned images are forwarded to the merging network, which generates the final HDR image.

The overall HDR imaging architecture is a cascade of an alignment network and a merging network, as shown in Fig. 3.2. The alignment network consists of two ESBNs, one for generating the under-exposure image with the structure of the reference and the other for over-exposure. Then, the aligned and reference images are stacked and

forwarded to the merging network, which consists of a one-step residual network. The alignment and merging networks can be trained separately, or the overall network can be trained end-to-end, where the latter yields better performance as expected.

Formally describing the process in Fig. 3.2, the overall network takes as input a set of three LDR images $\mathcal{I} = \{I_1, I_2, I_3\}$, and outputs an HDR image H. The input images I_1, I_2, and I_3 have low, medium, and high exposures, respectively, and they are presumed to have both background and foreground motions. This dissertation sets the image with medium exposure I_2 as the reference, and the ESBNs generate the aligned multi-exposure images $\{I_1^a$ or $I_3^a\}$ from $\{I_1$ or $I_3\}$ and I_2. As a result, the alignment network gives a set of aligned LDR images $\mathcal{I}^a = \{I_1^a, I_2^a, I_3^a\}$, where $I_2^a = I_2$, and the merging network generates H from $\mathcal{I}^a$.

3.2.2 Alignment Network

This dissertation proposes a CNN that generates the aligned image from the ones with foreground/background motions and also with a large difference in illumination. To synthesize an image with the illumination of I_s (namely the *source* image), $s = 1$ or 3, that have the same structure as I_r (*reference*), $r = 2$, the architectures shown in Fig. 3.3 can be considered: (1) Use only I_s as the input, and the CNN geometrically transforms the structure of I_s to the ground truth (that has the illumination of I_s and the structure of I_r). That is, this CNN is the geometric transformer. (2) Conversely, only I_r is input to the CNN, which changes its illumination to that of I_s so that it becomes close to the ground truth. The role of this CNN is just to change the illumination of I_r. (3) For increasing the information to be used in the CNN, both I_r and I_s are used as the input to the CNN. This network blends the images starting from the spatial domain and finally construct the image similar to the ground truth. (4) I_r and I_s go through different encoders so that their features are separately extracted, and they are blended in the feature domain and reconstructed to be close to the ground truth.

Fig. 3.3 also shows the result of each method, showing the closeness of the output

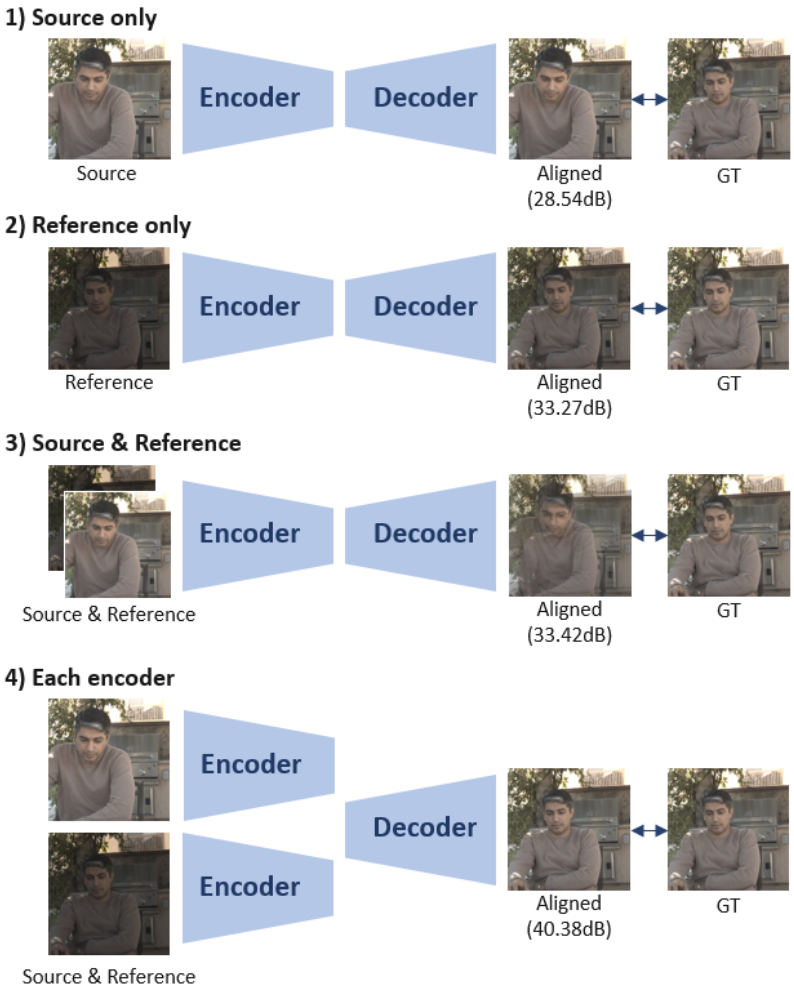


Figure 3.3: Four different architectures of CNN to generate the different exposure image aligned to the reference.

to the ground truth in terms of PSNR. As expected, using both I_r and I_s, and blending them in the feature domain synthesizes the best aligned image, which is the ESBN. The details of ESBN is shown in Fig. 3.4, which is inspired from the performances of *Unet* [65] and its advanced version, *Image Transformation Network* [32]. While they take a single image as the input of their encoder-decoder architecture and extract features using an encoder, the ESBN takes two images, I_s and I_r, and has two corresponding encoders to extract features from the images. Then, they are blended through several convolution layers and reconstructed by a decoder. Formally, the ESBN is described as a function

$$\hat{I_s^a} = f_{E_s}(I_s, I_r), \tag{3.1}$$

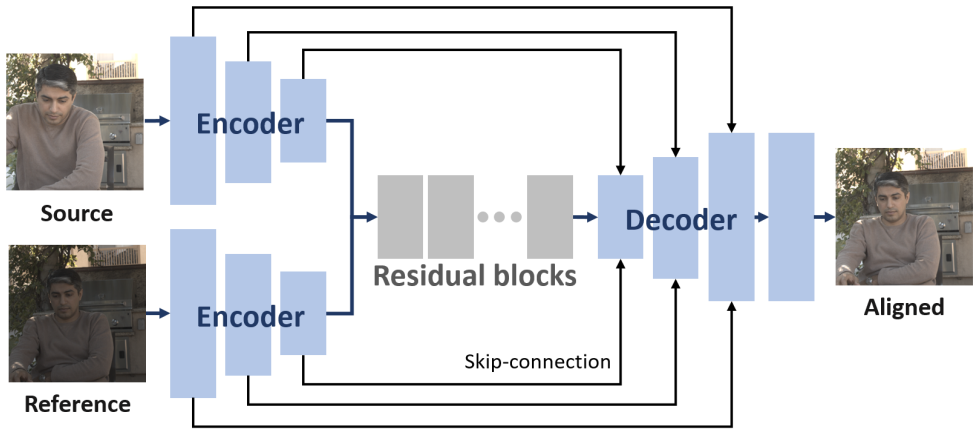
which means that the source I_s and the reference I_r are blended through the network f_{E_s} to generate the output $\hat{I_s^a}$.

To be more precise with the structure, each encoder has three encoding layers that are 5×5 convolution layers with the stride of 2, followed by batch normalizations and leaky ReLUs. The first layer of the encoder produces the feature map with 64 channels, and the second and third 128 and 256, respectively. The two encoders have their own parameters instead of sharing the same parameters. The feature maps from the upper and lower encoders are concatenated and then fed to the residual blocks [22]. There are nine residual blocks in total, each of which consists of 3×3 convolution layers, batch normalization, and a ReLU. Finally, the features are fed to the decoder to generate the aligned image. The decoder consists of three deconvolution layers with the stride of 1/2, which are followed by batch normalizations and ReLUs. Finally, there is an output layer, which is a 5×5 convolution layer with the stride of 1, followed by $\tanh(\cdot)$.

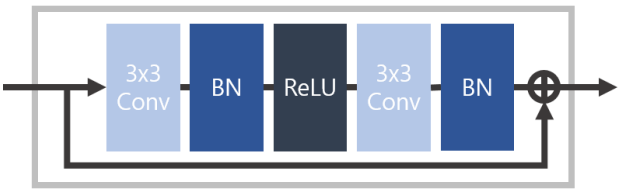
The ESBN $f_{E_s}(I_s, I_r)$ is trained by minimizing ℓ_2 distance between the aligned source and its ground truth, *i.e.,* by minimizing the loss function defined as

$$\mathcal{L}_{E_s} = \|\hat{I_s^a} - I_s^a\|_2, \tag{3.2}$$

where I_s^a is the ground truth for $\hat{I_s^a}$. ℓ_1 loss is also tried, but the ℓ_2 loss produces



(a) ESBN



(b) Residual block

Figure 3.4: (a) The details of ESBN. It consists of two encoders which extract the features from their input images, and residual blocks that blend the features, and a decoder that reconstructs the aligned image from the features. It also has skip-connections between the layers. (b) The details of the residual block in ESBN. It consists of two 3×3 convolution layers followed by batch normalizations and a ReLU.

quantitatively better aligned images (See Section. 3.4.2).

As shown in Fig. 3.2, two ESBNs, $f_{E_1}(I_1, I_2)$ for under-exposure and $f_{E_3}(I_3, I_2)$ for over-exposure, constitute the alignment network. The role of the alignment network is to produce under- and over-exposure images whose structures are aligned to that of the reference. Formally, the output from the alignment networks are summarized as

$$\hat{I_1^a} = f_{E_1}(I_1, I_2), \tag{3.3}$$

$$\hat{I_3^a} = f_{E_3}(I_3, I_2), \tag{3.4}$$

$$and \quad \hat{I}_2^a = I_2. \tag{3.5}$$

3.2.3 Merging Network

Recent HDR imaging methods [33, 78] reconstruct the final HDR image after aligning the input LDR images. But, since the aligned images still have artifacts or misalignments, they used somewhat complex networks to compensate for the misalignments while merging the images. On the other hand, since the ESBN works better than the optical flow or homography transformation, as will be demonstrated in the experiments, a simple network is used for the merging.

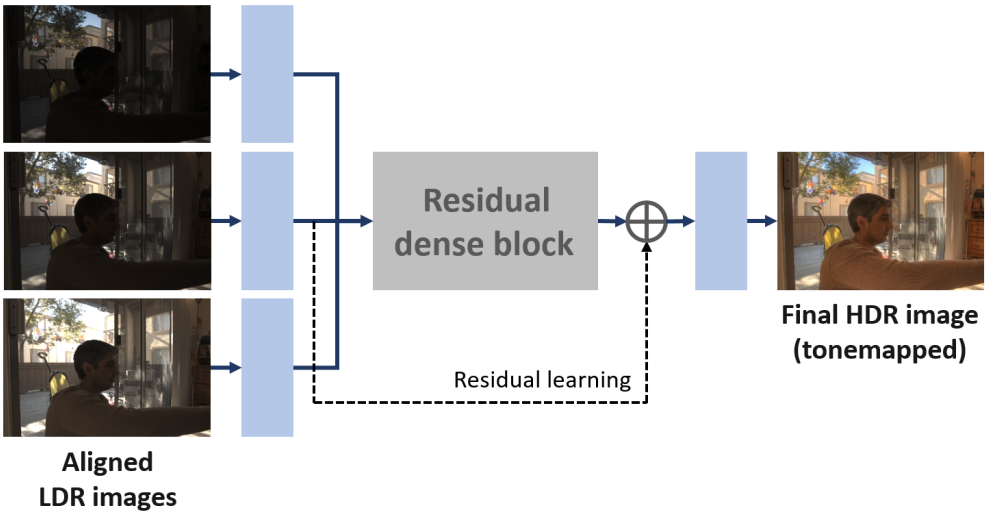


Figure 3.5: The architecture of the proposed merging network. The network extracts the features from the aligned LDR images and reconstructs the final HDR image. A residual dense block is used and a residual learning is applied to produce high-quality HDR images.

The proposed merging network extracts features from three aligned LDR images and reconstructs the final HDR images, as shown in Fig. 3.5. The residual learning with a residual dense block as in [43, 87] is adopted, while the original residual dense network includes three residual dense blocks. Each aligned image passes through 3×3

convolution layers and leaky ReLU, resulting in 64 feature maps. All the features are concatenated into 192 (64×3) channels, which are again reduced to 64 by using the 1×1 convolution layer. Then, it is fed to the residual dense block, which consists of six convolution layers followed by leaky ReLU. Finally, the HDR image is produced from the last output layer, which is a 3×3 convolution layer followed by $\tanh(\cdot)$.

For the residual learning, the output layer takes as input the sum of feature maps from the residual dense block and the reference. Formally, the merging network works as

$$\hat{H} = f_M(\mathcal{I}^a), \tag{3.6}$$

where $\hat{H}$ is the estimated HDR image and $\mathcal{I}^a$ is the set of aligned LDR images. The network is trained by minimizing ℓ_1 distance between the tonemapped estimated HDR image and the tonemapped ground truth HDR image, and the loss function $\mathcal{L}_M$ is defined as

$$\mathcal{L}_M = \|\mathcal{T}(\hat{H}) - \mathcal{T}(H)\|_1, \tag{3.7}$$

where H is the ground truth HDR image and $\mathcal{T}$ denotes the tonemapping operation. Since HDR images need to be tonemapped to the range of the targeting display, the loss function is computed in LDR domain instead of HDR domain as stated in [33]. A differentiable tonemapper, μ-law, is used, and it is defined as

$$\mathcal{T}(H) = \frac{\log(1 + \mu H)}{\log(1 + \mu)}, \tag{3.8}$$

where μ is set to 5,000.

3.2.4 Integrated HDR imaging network

The alignment network and merging network can be trained separately. However, it will be shown that end-to-end learning works better than the separate training, by minimizing the overall integrated loss

$$\mathcal{L}_I = \mathcal{L}_{E_1} + \mathcal{L}_{E_3} + \mathcal{L}_M. \tag{3.9}$$

3.3 Datasets

3.3.1 Kalantari Dataset and Ground Truth Aligned Images

For training the HDR imaging networks, multi-exposure images and ground truth HDR images are needed. There is a dataset with ground truth image for the static scene as in [17], but it is not easy to construct the dataset for the dynamic scenes that the camera and objects are moving. Recently, Kalantari *et al.* [33] made such a well-prepared dataset, where each set contains three dynamic exposure photos and a ground truth HDR image. Camera response functions of all multi-exposure LDR images in the dataset are calibrated using a gamma function. Hence, the proposed networks are trained with the Kalantari dataset. However, for the proposed method, ground truth aligned images I_1^a and I_3^a are also needed for training the alignment network. Thus, they are created from the ground truth HDR image. Specifically, the illumination of the ground truth HDR image is darken or lighten for each exposure by the gamma function used in [33].

3.3.2 Preprocessing

The ESBN takes six-channel images as the input, while general networks for image processing take three-channel images. In [33], they showed that the performance of their network could be improved by using both LDR images and linearized HDR ones as the input to the network. The reason for this result is that LDR images are suitable to extract features and to detect saturated regions, while HDR images contribute to improve the robustness of the network for the wide range of exposures. This dissertation follows this method for preparing the input to the network. Specifically, though Fig. 3.4 illustrates that the ESBN takes two images (over-exposed and medium-exposed), each image is actually the concatenation of LDR and its corresponding HDR images like [33].

3.3.3 Patch Generation

Since a large number of training examples are required to train the proposed networks, cropped patches instead of original images are used for the training. Specifically, 256×256 overlapping patches are cropped from 1500×1000 original images with a stride of 64, and around 18K patches are generated. Then, they are increased to 140K by data augmentation such as rotation and flipping. Furthermore, patches without background or foreground motions are removed, and the remaining 133K patches are finally used to train the networks efficiently.

3.4 Experimental Results

3.4.1 Evaluation Metrics

For the quantitative evaluation, three metrics, PSNR, SSIM, and HDR-VDP-2 [54] are used. Specifically, PSNR and SSIM are used to evaluate the aligned results, and HDR-VDP-2, PSNR, and SSIM are used to show the quality of the final HDR image. When computing PSNR and SSIM of HDR results, they are computed both in HDR domain (PSNR-L and SSIM-L) and LDR domain (PSNR-T and SSIM-T) after tonemapping.

3.4.2 Ablation Studies

Comparison of Loss Functions

Both ℓ_1 and ℓ_2 losses are tried for training the networks because they have pros and cons, which is not totally predictable. Hence, it is needed to choose one of them experimentally. In general, the ℓ_1 is more robust to outliers than the ℓ_2, because the outliners make the squared values of differences (ℓ_2) too large. On the other hand, the ℓ_2 is known to work more stable when pairs of training images do not have significant differences [66].

First, the performances of the proposed alignment network and merging network trained with different loss functions are compared. The alignment network is trained using ℓ_1 and ℓ_2 loss, and the alignment results on Kalantari test sets [33] are listed in Table. 3.1. The results indicate that the alignment network trained using ℓ_2 yields better performance for the alignment. Similarly, the merging network is trained using ℓ_1 and ℓ_2, and the reconstructed HDR results are compared in Table. 3.2. It can be seen that ℓ_1 shows better performance in the case of the merging network.

Comparison of Separate and Joint Training

In Section. 3.2.4, this dissertation explained that the alignment network and merging network can be trained separately, or jointly trained as a single network. In this subsec-

Table 3.1: Comparisons of the alignment network trained with different loss functions. The best results are shown in boldface.

	PSNR	SSIM
ℓ_1 loss	39.5832	0.9815
ℓ_2 loss	**40.3849**	**0.9880**

Table 3.2: Comparison of the merging network trained with different loss functions. The best results are shown in boldface.

	PSNR-T	SSIM-T	PSNR-L	SSIM-L
ℓ_1 loss	**44.0313**	**0.9914**	**41.1796**	**0.9871**
ℓ_2 loss	43.6762	0.9860	40.7847	0.9845

tion, the results of these training methods are compared in Table. 3.3, which shows that the joint training brings performance gains in HDR reconstruction. However, the network trained jointly without the loss for the ESBN has lower alignment performance, while the network trained jointly with the loss for the ESBN maintains the alignment performance.

Table 3.3: Comparison of the separate and joint training of the proposed network.

	Alignment		HDR	
	PSNR	SSIM	PSNR	SSIM
Separate	40.38	0.9880	44.03	0.9914
Joint (w/o ESBN loss)	39.06	0.9796	44.56	0.9928
Joint (w/ ESBN loss)	40.25	0.9865	44.55	0.9930

3.4.3 Comparisons with State-of-the-Art Methods

Image Alignment

This dissertation first compares the alignment accuracy by ESBN and other alignment methods employed in previous HDR imaging methods on the test sets in Kalantari dataset. In Fig. 3.6, the source image with high exposure is aligned to the reference using the proposed and other alignment methods. It can be seen that the method of Hu *et al.* [25] using patch matching reconstructs the aligned image without artifacts, but it has low contrast and loses details. The method of Kalantari *et al.* [33] using optical flow failed to estimate motions and produces some artifacts in the foreground regions. The method of Wu *et al.* [78] using homography transformation does not align the foreground motions, while it aligns the background. Our alignment network generates the aligned image very close to the ground truth image. PSNR and SSIM of the aligned results are also computed in Table. 3.4 to compare the performance quantitatively. It can be seen that the proposed alignment method shows better performance than the others.

(a) Input (b) Aligned result

(c) Hu (d) Kalantari (e) Wu (f) Proposed (g) GT

Figure 3.6: Comparsions of the proposed ESBN and other alignment methods. (a) A source image and a reference image. (b) The aligned image using the ESBN. Magnification of red and green boxes by (c) Hu *et al.* [25], (d) Kalantari *et al.* [33], (e) Wu *et al.* [78], (f) proposed method, and (g) the ground truth.

HDR Reconstruction

This dissertation compare the HDR imaging results of the proposed method with several state-of-the-art methods on test images in Kalantari dataset. In Fig. 3.7, the reconstructed HDR images of various methods are visualized, which are tonemapped with Photomatix to be visualized on the LDR displays. The patch-based reconstruc-

Table 3.4: Comparisons of alignment results by ESBN and other alignment methods. The best results are shown in boldface.

	Hu [25]	Kalantari [33]	Wu [78]	Proposed
PSNR	34.4117	32.5710	28.3794	**40.3849**
SSIM	0.9683	0.9351	0.8614	**0.9880**

tion methods by Sen *et al.* [68] and Hu *et al.* [25] do not well reconstruct the saturated regions on the branches and the wall. They fail to find correct corresponding patches in the saturated regions and reconstruct the regions with the patches in the sky. The CNN-based methods of Kalantari *et al.* [33] and Wu *et al.* [78] reconstruct the saturated regions in some degree, but they fail to reconstruct the details of the shadow of the tree on the wall. On the contrary, the proposed HDR imaging network can reconstruct both the saturated regions and the details. The results of Yan *et al.* [79] are not shown here because the source code is not yet provided.

In Table. 3.5, PSNR, SSIM and HDR-VDP-2 between the HDR results and the corresponding ground truth images are also computed to evaluate the methods quantitatively. It can be seen that the proposed network with separate training of the alignment and merging network provides better performance than the others. Moreover, the joint training of the network brings a performance gain. This dissertation quote the quantitative evaluation results of Yan *et al.* [79] from their paper, since their source code is not available.

The running times are also compared in Table. 3.6, by executing the source codes[1] provided by the authors. The codes were run on a PC with i7-4790 and Titian XP, for the LDR images of 1500×1000. The execution time for image alignment is measured separately, except for Sen and Yan that directly reconstruct the HDR image. It can be

[1] Since the source code of the alignment using homography transformation in [78] is not provided, the alignment code which can generate similar results is implemented directly.

(a) Input (b) HDR result

(c) Sen (d) Hu (e) Kalantari (f) Wu (g) Proposed (h) GT

Figure 3.7: Comparsions of proposed and recent HDR imaging methods. (a) A set of input LDR images. (b) HDR result of the proposed method. Magnification of red and green boxes by (c) Sen *et al.* [68], (d) Hu *et al.* [25], (e) Kalantari *et al.* [33], (f) Wu *et al.* [78], (g) proposed method, and (h) the ground truth.

seen that the alignment process has the shortest running time because it is performed by the GPU, while other alignment methods are performed by the CPU. Furthermore, the merging network takes less time than those of Kalantari *et al.* [33] and Wu *et al.* [78] because the network has a simple structure.

Table 3.5: Comparisons of HDR results of the proposed method with state-of-the-art methods. The best results are shown in boldface.

	PSNR-T	SSIM-T	PSNR-L	SSIM-L	HDR-VDP-2
Sen [68]	41.11	0.9815	38.82	0.9749	59.38
Hu [25]	34.87	0.9701	31.72	0.9520	57.05
Kalantari [33]	42.70	0.9878	41.21	0.9845	62.98
Wu [78]	41.65	0.9857	40.88	0.9849	62.70
Yan [79]	43.62	-	41.03	-	62.30
Proposed (separate)	44.03	0.9914	41.18	0.9871	63.02
Proposed (joint)	**44.55**	**0.9930**	**41.36**	**0.9877**	**63.24**

3.4.4 Application to the Case of More Numbers of Exposures

Note that Fig. 3.2 shows the case of three-exposures ($N = 3$), but it is believed that the proposed method can straightforwardly extend this structure for $N > 3$. However, end-to-end learning for $N > 3$ is currently not possible, like all the recent deep-learning-based dynamic scene HDR imaging methods that use Kalantari dataset for the training [33, 78, 79]. For the extension, it is needed to make a well-prepared dataset with more numbers of exposures, which would be much more complicated than the three-exposure case. Hence, very high-cost for the dataset construction is the limitation of the proposed and existing methods.

But, it is shown that the ESBN, which is separately trained, can be applied for the merge of a larger number of exposures. Precisely, the network generates aligned images with significant exposure differences very well, and hence can be used as a pre-processing step for any conventional HDR imaging methods for $N > 3$. For example, the ESBNs are applied to an image set with $N = 5$ in Sen dataset [68], and the alignment results are shown in Fig. 3.8. It can be seen that the ESBNs generate the aligned images very well, even though the exposure differece is very large. Note that

Table 3.6: Comparisons of average execution times of the proposed method with state-of-the-arts.

	Alignment(s)	Merging(s)	Total(s)
Sen [68]	-	-	312.853
Hu [25]	185.722	1.712	187.434
Kalantari [33]	29.893	1.558	31.451
Wu [78]	0.501	0.239	0.740
Yan [79]	-	-	0.320
Proposed	0.231	0.172	0.403

the ESBN is not retrained for the Sen dataset (the ESBN trained with the Kalantari dataset is applied in this case), which shows its generalization performance. In Fig. 3.8, the HDR results using the original version of the HDR imaging method in [25] and the modified version where the alignment method is substituted with the proposed alignment method are compared. The latter reconstructs more natural sky, while some artifacts are shown in the former. In summary, it is believed that the ESBN is also effective for the case of $N > 3$, and can be used as a preprocessing step for other merging methods as well.

(a) Input images

(b) Aligned images

(c) Hu (d) Ours

Figure 3.8: (a) An image set with $N = 5$ which contains images with significant exposure differences in Sen dataset. (b) Alignment results using the ESBNs. The comparison of HDR results by (c) Hu *et al.* [25], where they used the merging method of [57] and (d) the proposed alignment method, also with merging by [57].

3.4.5 Pre-processing for other HDR imaging methods

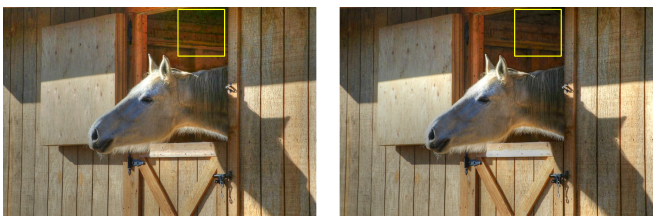
To show the effectiveness of the proposed ESBN as a preprocessor for other HDR imaging methods, the HDR results by the original version of the state-of-the-art methods and the modified versions which include the proposed ESBN is compared. Precisely, the aligned multi-exposure images are first generated by the ESBN and then the final HDR image is constrcuted by the original HDR imaging method. In Fig. 3.9, the HDR results of an image set with $N = 3$ are compared. The methods of Sen *et al.* [68] and Kalantari *et al.* [33] produce some artifacts, but those disappear in the modified version of each method. The method of Hu *et al.* [25] fails to reconstruct the dark region, but the modified method succeeds. The modified method of Wu *et al.* [78] reconstructs the mane of the horse with more details. In addition, the HDR results of an image set with $N = 5$ are compared in Fig. 3.10. The modified method of Sen *et al.* [68] reconstructs the yellow cap of the pen and the highlighted side of the clock more naturally than before, and the modified method of Hu *et al.* [25] causes less artifact.

PSNR and SSIM of the HDR results of test images in Kalantari dataset [33] for the modified methods are computed. Table. 3.7 shows that the state-of-the-art HDR imaging methods are quantitatively improved, and especially the method of Hu *et al.* [25] is largely improved by using the ESBN as a preprocessing step.

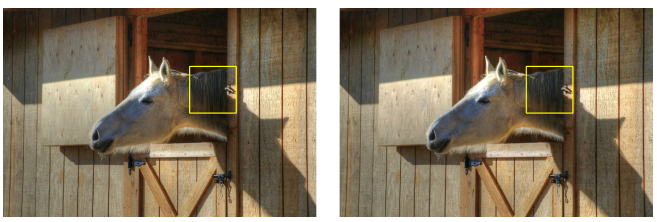
(a) Input images

(b) Sen [68] and ESBN+Sen

(c) Hu [25] and ESBN+Hu



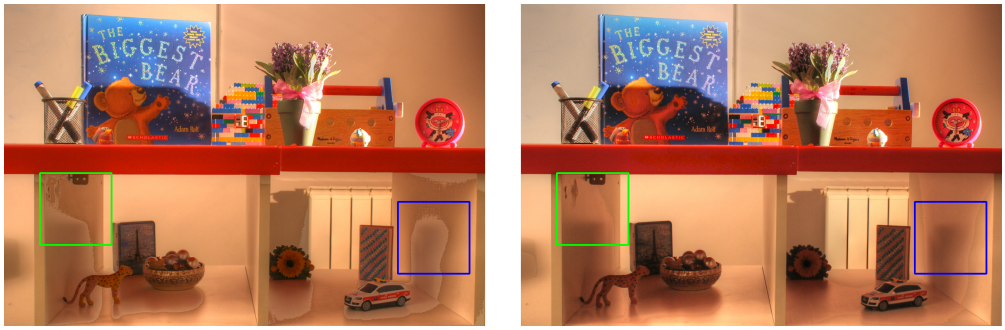
(d) Kalantari [33] and ESBN+Kalantari



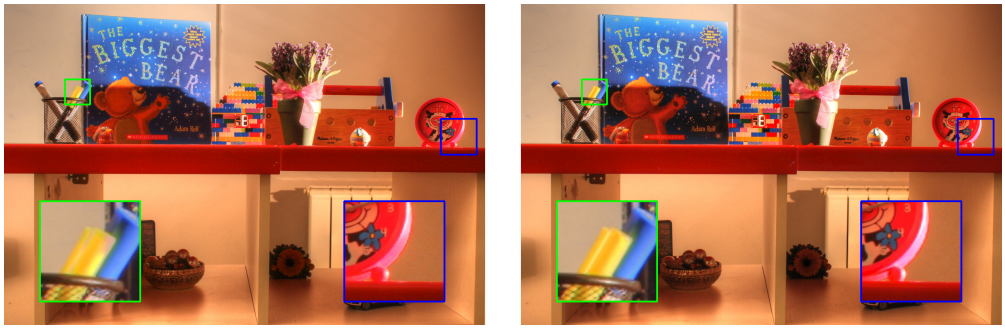
(e) Wu [78] and ESBN+Wu

Figure 3.9: (a) An input image set with $N = 3$ by courtesy of Sing Bing Kang [35]. (b) (c) (d) (e) Comparisons of the HDR results by the original and the modified version of the state-of-the-art methods.

(a) Input images



(b) Sen [68] and ESBN+Sen



(c) Hu [25] and ESBN+Hu

Figure 3.10: (a) An input image set with $N = 5$ by courtesy of Karaduzovic-Hadziabdic [37].. (b) (c) Comparisons of the HDR results by the original and the modified version of the state-of-the-art methods.

Table 3.7: Comparisons of the original version of the state-of-the-art HDR imaging methods and the modified version where the ESBN is used as a preprocessing. The better results are shown in boldface.

	PSNR-T	SSIM-T	PSNR-L	SSIM-L
Sen [68]	41.11	0.9815	38.82	0.9749
ESBN+Sen [68]	**41.35**	**0.9863**	**39.70**	**0.9800**
Hu [25]	34.87	0.9699	31.72	0.9520
ESBN+Hu [25]	**38.99**	**0.9870**	**36.31**	**0.9808**
Kalantari [33]	42.70	0.9877	**41.21**	0.9845
ESBN+Kalantari [33]	**43.03**	**0.9908**	41.11	**0.9860**
Wu [78]	**41.65**	0.9860	40.88	**0.9858**
ESBN+Wu [78]	41.60	**0.9890**	**41.44**	0.9856

Chapter 4

Video Super-resolution

4.1 Motivation

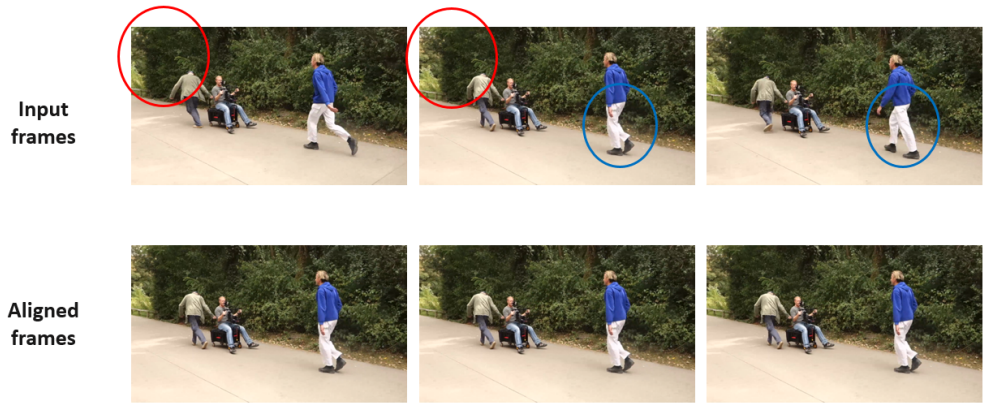


Figure 4.1: The input LDR images before which have both background (red circles) and foreground motions (blue circles), and the aligned result.

In practical situations, there are misalignments in the consecutive frames in videoes, due to background and foreground motions as shown in Fig. 4.1. Before directly merging the frames to reconstruct the HR frame, the input images are should be aligned. Recent methods first align the frames to the reference target frame by adding spatial transformer network (STN) to their overall VSR network [6, 21, 47, 67]. However,

their aligned results usually have some artifacts in regions with complex background or non-rigid foreground motions. Thus, they treat the artrifacts in the HR reconstruction step, or some methods don't conduct the explicit motion compensation [30, 84]. Thus, this dissertation propose an alignment method using a CNN which is modified version of the ESBN to generate well-aligned frames without artifacts.

4.2 Proposed Method

4.2.1 Overall Pipeline

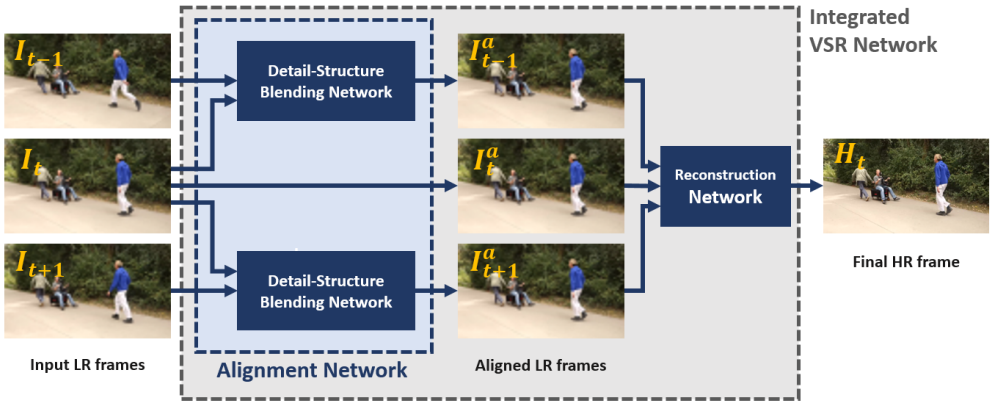


Figure 4.2: The overall pipeline of the proposed VSR method for the case of $N = 3$. First, structurally aligned frames are generated in the Alignment Network, which consists of detail-structure blending networks (DSBNs). Then, the aligned images are forwarded to the reconstruction network, which generates the final HR frames.

The overall video super-resolution (VSR) architecture is a cascade of an alignment network and a reconstruction network, as shown in Fig. 4.2. For the N input frames, the alignment network consists of $N-1$ DSBNs, each for generating the corresponding neighbor frames with the structure of the reference. Then, the aligned and reference frames are concatenated and forwarded to the reconstruction network, which consists of several residual blocks and a decoder. The alignment and reconstruction networks

can be trained separately, or the overall network can be trained end-to-end. As shown in Sec. 3.4.2, the latter is expected to yields better performance.

Formally describing the process in Fig. 4.2, the proposed overall network takes as input a set of 7 consecutive LR frames $\mathcal{I}_t = \{I_{t-3}, I_{t-2}, ..., I_{t+3}\}$, and outputs an HR frame H_t. The set consists of the target frame I_t, the past neighbor frames $\{I_{t-3}, I_{t-2}, I_{t-1}\}$ and the future neighbor frames $\{I_{t+1}, I_{t+2}, I_{t+3}\}$, and they are presumed to have both background and foreground motions. the target frame I_t is set as the reference, and the DSBNs generate the aligned neighbor frames I^a_{t+k} from I_{t+k} and I_t. As a result, the alignment network gives a set of aligned LR frames $\mathcal{I}^a_t = \{I^a_{t-3}, I^a_{t-2}, ..., I^a_{t+3}\}$, where $I^a_t = I_t$, and the reconstruction network generates H_t from $\mathcal{I}^a_t$.

4.2.2 Alignment Network

The proposed ESBN is first used intactly to the image alignment process in VSR. It takes two images, the source image I_s and the reference I_r, and has two corresponding encoders to extract features from the images. Then, they are blended through several convolution layers and reconstructed by a decoder. Formally, the network is described as a function

$$\hat{I^a_s} = f_D(I_s, I_r), \tag{4.1}$$

which means that the source I_s and the reference I_r are blended through the network f_D to generate the output $\hat{I^a_s}$. It is renamed as detail-structure blending network (DSBN), since it is designed to blend the detail of I_s and the structure of I_r while the ESBN is designed to blend the exposedness of I_s and the structure of I_r. The network $f_D(I_s, I_r)$ is trained by minimizing ℓ_1 distance between the aligned source and its ground truth, *i.e.,* by minimizing the loss function defined as

$$\mathcal{L}_D = \|\hat{I^a_s} - I^a_s\|_1, \tag{4.2}$$

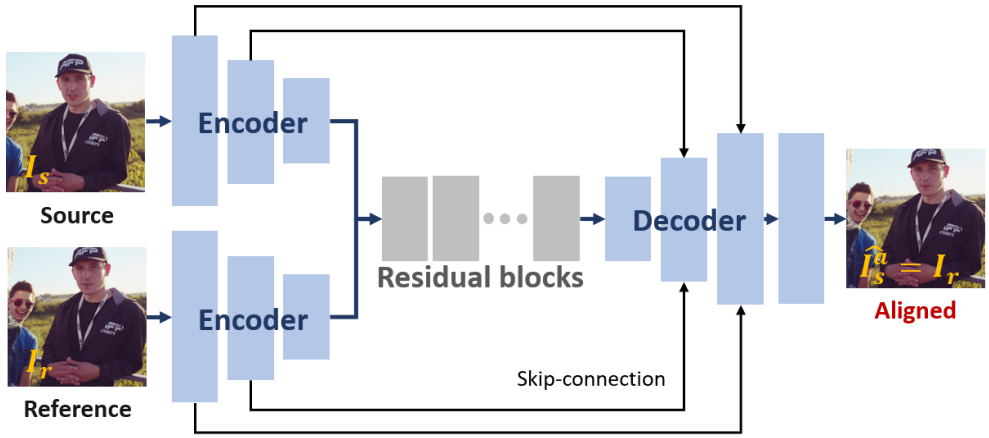


Figure 4.3: The details of DSBN. It consists of two encoders which extract the features from their input images, and residual blocks that blend the features, and a decoder that reconstructs the aligned image from the features. It also has skip-connections between the layers.

where I_s^a is the ground truth for $\hat{I_s^a}$, and the reference I_r can be used as the ground truth since it doesn't exist actually.

However, the DSBN is trained to have a trivial solution in this case, which means that the network can reconstruct the output only with I_r and without I_s. To deal with this problem, the reference is blurred before forwarded to the corresponding encoder. Then, it is hard for the network to recosntruct $\hat{I_s^a}$ only with blurred I_r, and it uses information from I_s as well as I_r. The blurring is performed by using a 3×3 median filter. The effect of this tricky method is proved by comparing two experimental results in Sec. 4.3.3.

To be more precise with the structure, each encoder has three encoding layers that are 3×3 convolution layers with the stride of 2, followed by leaky ReLUs. The first layer of the encoder produces the feature map with 64 channels, and the second and third 128 and 256, respectively. The two encoders have their own parameters instead of sharing the same parameters. The feature maps from the upper and lower encoders are concatenated and then fed to the residual blocks [22]. There are five residual blocks in

total, each of which consists of 3×3 convolution layer and a ReLU. The features are fed to the decoder to generate the aligned image, which consists of three deconvolution layers with the stride of 1/2, which are followed by leaky ReLUs. Finally, there is an output layer, which is a 3×3 convolution layer with the stride of 1, followed by $\tanh(\cdot)$.

As shown in Fig. 4.2, $N-1$ DSBNs for $N-1$ neighbor frames, constitute the alignment network. The role of the alignment network is to generate $N-1$ neighbor frames, whose structures are aligned to that of the reference. Formally, the output from the alignment networks are summarized as

$$\hat{I^a_{t+k}} = f_{D_k}(I_{t+k}, I_t) \tag{4.3}$$

$$and \quad \hat{I^a_t} = I_t, \tag{4.4}$$

where $k = -3, -2, -1, +1, +2, +3$

4.2.3 Reconstruction Network

Recent VSR methods [6,67] reconstruct the final HR frame after aligning the input LR frames. But, since the aligned frames still have artifacts or misalignments, they used somewhat complex networks to compensate for the misalignments while merging the frames. On the other hand, since the proposed DSBN works better than other alignment methods in the existing VSR methods, a simple network is used for the merging and reconstrcution.

The proposed reconstruction network extracts features from the concatenation of aligned LR frames and reconstructs the final HR frame, as shown in Fig. 3.5. The network consists of 5 residual blocks and a decoder. All aligned frames are concatenated and pass through a 3×3 convolution layer and a leaky ReLU, resulting in 128 feature maps. Then, it is fed to the residual blocks, each of which consists of a 3×3 convolution layer followed by a leaky ReLU and another 3×3 convolution layer. Finally, the feature maps are fed to the decoder to generate the HR frame. The decoder consists of

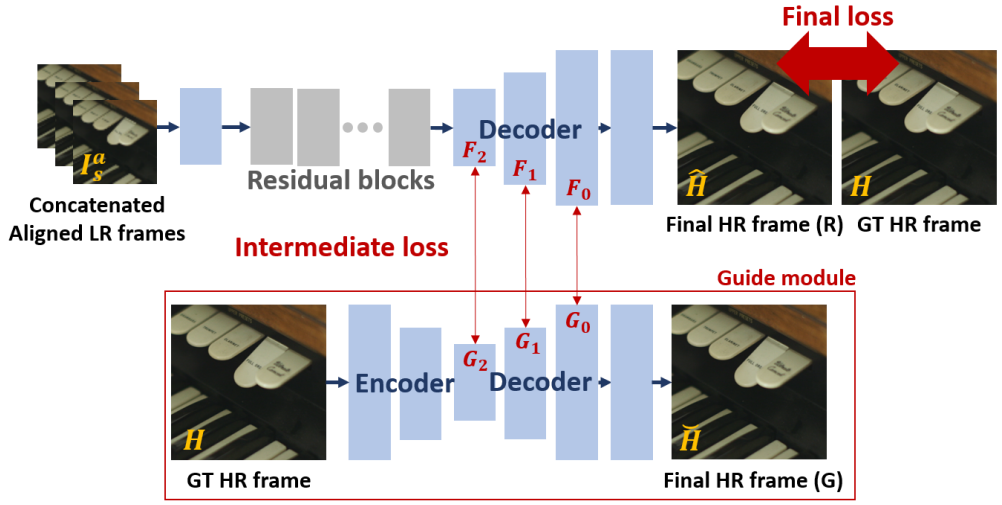


Figure 4.4: The architecture of the proposed reconstruction network. The network extracts the features from the aligned LR frames and reconstructs the final HR frame. Several residual blocks and a decoder is used to generate high-quality HR frames.

three deconvolution layers with the stride of 1/2, which are followed by leaky ReLUs. Finally, there is an output layer, which is a 3×3 convolution layer with the stride of 1, followed by $\tanh(\cdot)$.

Formally, the reconstruction network works as

$$\hat{H}_t = f_R(\mathcal{I}_t^a), \tag{4.5}$$

where $\hat{H}_t$ is the estimated HR frame and $\mathcal{I}_t^a$ is the set of aligned LR frames. The network is trained by minimizing ℓ_1 distance between the estimated HR frame and the ground truth HR frame, and the loss function $\mathcal{L}_R$ is defined as

$$\mathcal{L}_R = \|\hat{H}_t - H_t\|_1, \tag{4.6}$$

where H_t is the ground truth HR frame.

4.2.4 Integrated VSR network

The alignment network and reconstruction network can be trained separately. However, it will be shown that end-to-end learning works better than the separate training, by minimizing the overall integrated loss

$$\mathcal{L}_I = \sum_{k=-3}^{3} \mathcal{L}_{D_k} + \mathcal{L}_R. \tag{4.7}$$

4.3 Experimental Results

4.3.1 Dataset

For training the proposed networks, Vimeo-90k septuplet dataset is used. Vimeo-90k septuplet dataset is a dataset for video super-resolution, and the subset of Vimeo-90k dataset which include 90k video clips from Vimeo. It consists of 64,612 septuplets from 39k video clips, and each septuplet contain 7 consecutive frames whose resolution is 448×256. 256×256 patches are randomly cropped from original frames, and they are increased to 260K by data augmentation such as rotation and flipping. These HR patches are resized to 64×64 LR patches and used as input of the proposed networks. For test, Vid4 benchmark dataset and YT10 dataset are used.

4.3.2 Ablation Study

Comparison of Model Architectures

The architecture of the proposed integrated VSR network is investigated and the importance of different individual modules in the whole network is validated. This ablation study is achieved by comparing the proposed VSR network and the following variants.

- **SR :** For the VSR baseline without any motion compensation process, the alignmnet network is omitted from the proposed VSR network. The concatenation of input LR frames $\mathcal{I}_t$ is directly feeded into the reconstruction network.

- **STN-SR :** For the VSR baseline with an existing motion compensation process, the DSBNs of the proposed VSR network is replaced to the STN based motion compensation modules in [67]. A set of input LR frames $\mathcal{I}_t$ is aligned using the modules and feeded into the reconstruction network.

- **DSBN-SR :** The proposed alignment network which consists of DSBNs and reconstruction network is used. However, the reconstruction network is trained without the guide module. A set of aligned LR frames $\mathcal{I}_t^a$ is generated using the DSBNs and feeded into the reconstruction network.

- **DSBN-SR-GM :** The proposed integrated VSR network is used A set of aligned LR frames $\mathcal{I}_t^a$ is generated using the DSBNs and feeded into the reconstruction network trained with the guide module.

In Table. 4.1, each component of the proposed VSR network is validated for performance gains.

Table 4.1: Comparisons of the proposed VSR network with different model architectures. The best results are shown in boldface.

	PSNR	SSIM
SR	25.06	0.7194
STN-SR	25.38	0.7579
DSBN-SR	26.79	0.8241
DSBN-SR-GM	**27.01**	**0.8297**

Comparison of Separate and Joint Training

In Section. 4.2.4, it is explained that the proposed alignment network and reconstruction network can be trained separately, or jointly trained as a single network. In this

subsection, the results of these training methods are compared. Table. 4.2 shows that the joint training brings performance gains.

Table 4.2: Comparisons of the separate and joint training of the proposed VSR network. The best results are shown in boldface.

	PSNR	SSIM
Separate	27.01	0.8297
Joint	**27.15**	**0.8322**

Table 4.3: Comparisons of the different loss functions for the HR reconstruction network with guide module. The best results are shown in boldface.

	PSNR	SSIM
ℓ_{final}	26.79	0.8241
$\ell_2 + \ell_{final}$	27.01	0.8297
$\ell_2 + \ell_1 + \ell_0 + \ell_{final}$	27.03	0.8291
$\ell_2 + \frac{1}{2}\ell_1 + \frac{1}{4}\ell_0 + \frac{1}{4}\ell_{final}$	**27.17**	**0.8328**

4.3.3 Capability of DSBN for alignment

In Sec. 4.2.2, a tricky method which is to blur the reference I_r is used to let the DSBN do its proper role. To prove that the DSBN generate the aligned frames I_s^a with information from both the source I_s and the reference I_r, two experiments are designed. As shown in Fig 4.5, the source and the blurred reference are used as inputs of the DSBN in the first case, and the face of the baby in the source is covered with a white rectangle in the second case. The DSBN generates the aligned frames well in the first case, while it generates the unsuccessful result in the second case. Thus, this experi-

ments shows that the DSBN generate the aligned result by blending the information from both images not deblur the reference image. Furthermore, it also has quantitatively better performaces than the STN based alignment method in [67] as shown in Table. 4.1.

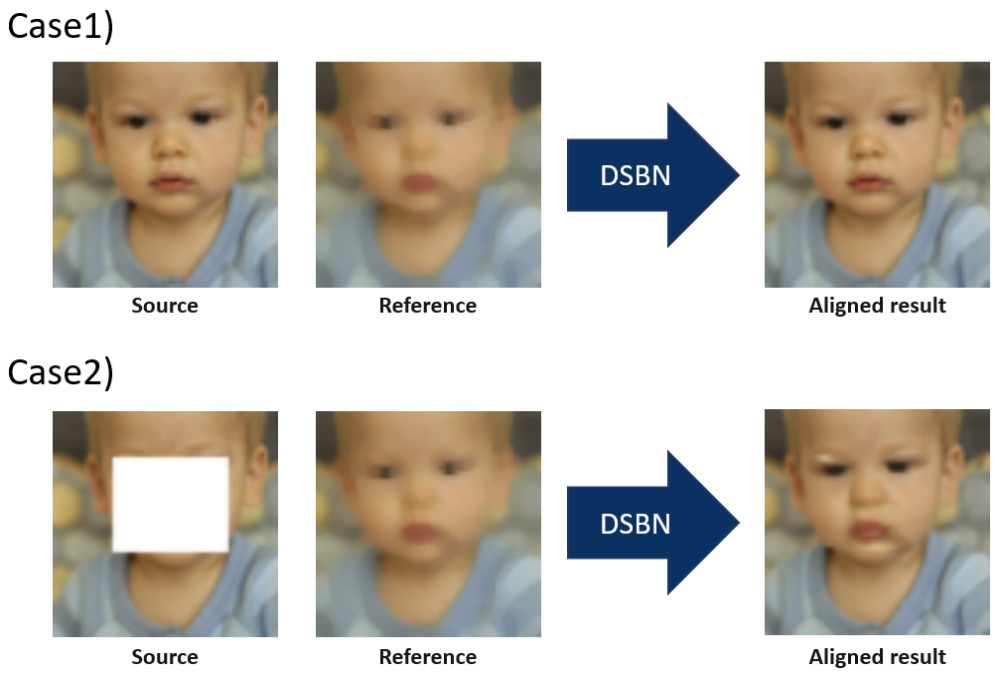


Figure 4.5: Two cases of experiments to prove the capability of the DSBN. (Case1) The source and the blurred reference are used as inputs of the DSBN. (Case2) The source where the region is covered with a white rectangle and the blurred reference are used.

4.3.4 Comparisons with State-of-the-Art Methods

Video Super-resolution

The VSR results of the proposed method with the state-of-the-art methods including VESPCN [6], MCRes [47], FRVSR [67], RBPN [21], DUF [30] and PFNL [84] are first compared on Vid4 benchmark dataset. Some previous VSR methods generate in-

put LR frames by down-sampling HR frames bicubically, while recent methods apply Gaussian filters and then down-sample. Different down-sampling processes can affect the LR-to-HR mapping relationship the network tries to learn, thus, it is unfair to compare the methods under different down-sampling processes. In order to make a fair comparison with the state-of-the-art methods, the methods [6,30,47,67,84] whose networks are retrained with the same down-sampling process and same training dataset in [84] are used, and the proposed network and RBPN [21] is also trained under the same condition.

In Fig. 4.6, the reconstructed HR frames of the state-of-the-art methods are visualized. FRVSR [67] and RBPN [21] reconstruct horizontal pattern of the building better than VESPCN [6]. DUF [30] and PFNL [84] reconstruct vertical pattern as well as horizontal one, however fail to reconstruct windows. On the contrary, the proposed VSR method can reconstruct the windows in the buinlding.

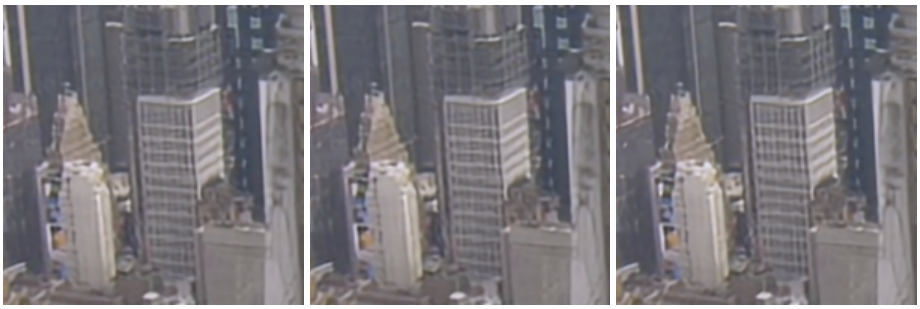
In Table. 4.4, the PSNR and SSIM between the VSR results and the corresponding ground truth frames are computed to evaluate the methods quantitatively. Vid4 dataset only contains scenes with low resolution (approximately 720×540), YT10 dataset which includes scenes with higher resolution (approximately 1280×720) is additionally used. The PSNR and SSIM are calculated only on Y channel of YCbCr colorspace for all frames except the first and last two frames. The proposed method doesn't achieve the best performance, but is close to PFNL [84] which has the best performance.

The execution times are also compared in Fig. 4.8, by executing the source codes provided by [21, 84]. The codes were run on a PC with i7-4790 and Titian XP, for the scenes in Vid4 dataset. The proposed method is the fastest among the 4 methods which are close to the best performance, and especially twice faster than PFNL [84] which has the best performance.

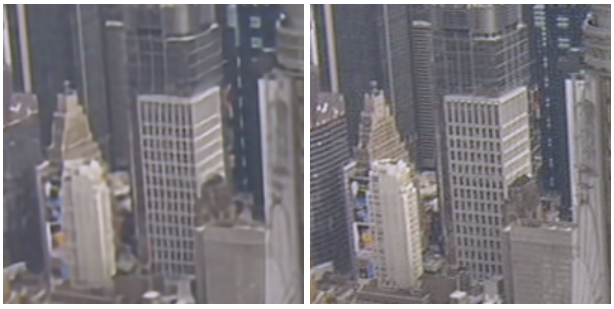
Table 4.4: Comparisons of VSR results of the proposed method with state-of-the-art methods. The best results are shown in boldface.

	Vid4		YT10	
	PSNR	SSIM	PSNR	SSIM
VESPCN [6]	25.52	0.7510	35.62	0.9483
MCRes [47]	25.81	0.7676	35.90	0.9424
FRVSR [67]	26.69	0.8105	37.16	0.9501
RBPN [21]	27.12	0.818	37.95	0.9572
DUF [30]	27.18	0.8265	38.06	0.9585
PFNL [84]	**27.40**	**0.8384**	**38.93**	**0.9641**
Proposed	27.15	0.8322	38.18	0.9603

(a) Input (b) VESPCN [6] (c) FRVSR [67]



(d) RBPN [21] (e) DUF [30] (f) PFNL [84]



(g) Proposed (h) GroundTruth

Figure 4.6: Comparsions of the proposed and state-of-the-art method VSR methods. (a) Input target frame, (b) VESPCN [6], (c) FRVSR [67], (d) RBPN [21], (e) DUF [30], (f) PFNL [84], (g) Proposed method and (h) Ground Truth.

(a) Input (b) VESPCN [6] (c) FRVSR [67]

(d) RBPN [21] (e) DUF [30] (f) PFNL [84]

(g) Proposed (h) GroundTruth

Figure 4.7: Comparsions of the proposed and state-of-the-art method VSR methods. (a) Input target frame, (b) VESPCN [6], (c) FRVSR [67], (d) RBPN [21], (e) DUF [30], (f) PFNL [84], (g) Proposed method and (h) Ground Truth.

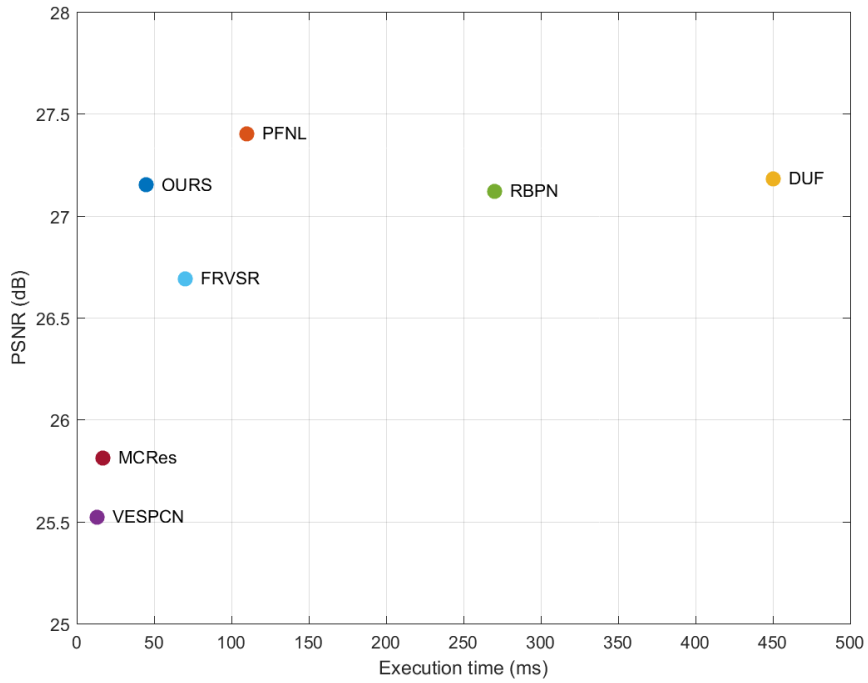


Figure 4.8: Comparisons of execution time over performance of the proposed method with state-of-the-art methods.

Chapter 5

Joint HDR and SR

5.1 Proposed Method

5.1.1 Feature Blending Network

In Chapter. 3 and Chapter. 4, it is shown that two image fusion tasks, high dynamic range (HDR) imaging and video super-resolution (VSR), are well-performed due to the proposed blending networks. Meanwhile, it is expected that dynamic range and resolution of the image can be enhanced simultaneously by fusing multiple images. For the enhancement, it is necessary for the blending network to synthesize the image with the exposedness and the detail of the source and the structure of the reference. The synthesis is performed by using the ESBN or DSBN and applying the trick used for the DSBN which is to blur the reference image. The new network for this joint HDR-SR task is renamed as feature blending network (FBN).

5.1.2 Joint HDR-SR Network

For the joint HDR-SR task, the proposed VSR network in Chapter. 4 can be used. The network architecture is unchanged, however the input and output of the network are changed as shown in Fig. 5.1. The network is trained to generate the final high

resolution (HR) and high dynamic range (HDR) image with three low resolution (LR) and low dynamic range (LDR) images, while it is trained to generate the HR image with seven LR images in the VSR task.

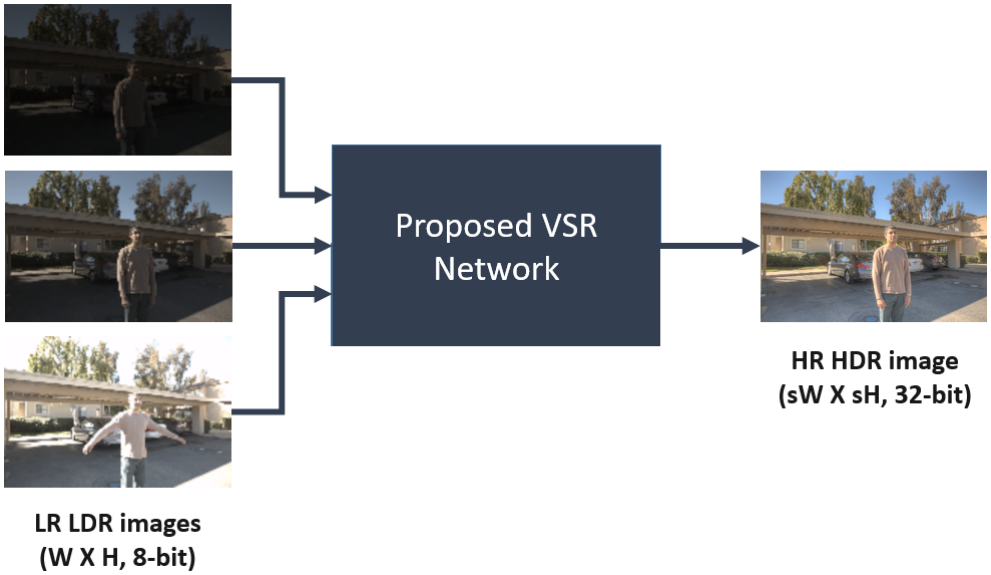


Figure 5.1: The joint HDR-SR scheme using the proposed network for VSR. LR and LDR images are fused into an HR and HDR image.

5.1.3 Existing VSR Network

The existing VSR network can be used likewise for the joint HDR-SR. Among the various VSR methods, VESPCN [6] and RBPN [21], which performs and doesn't perform the explicit motion compensation respectively are selected and trained for joint HDR-SR task. The architecture of the networks is unchanged, however the input and output of the networks are changed as shown in Fig. 5.2. The network is trained to generate the final HR and HDR image with three LR and LDR images, likewise in Sec. 5.1.2.

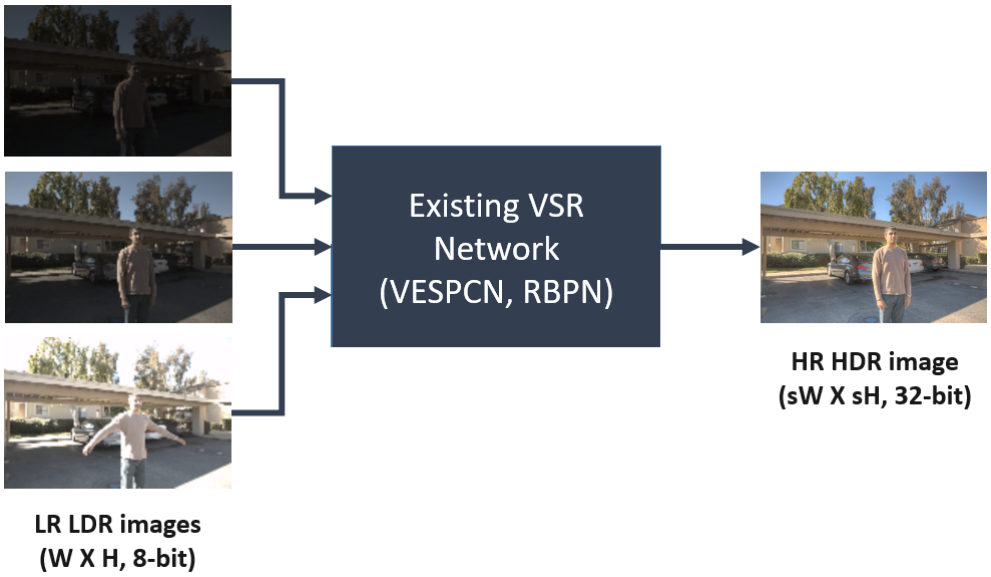


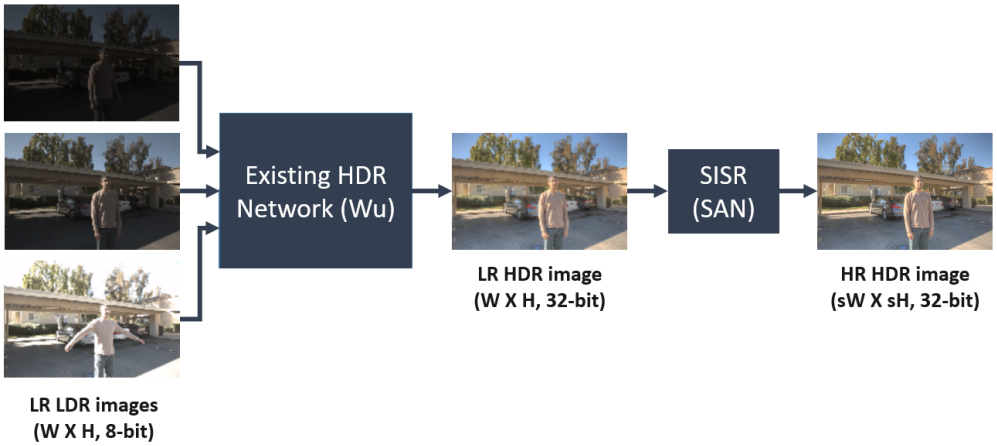
Figure 5.2: The joint HDR-SR scheme using the existing VSR network. LR and LDR images are fused into an HR and HDR image.

5.1.4 Existing HDR Network

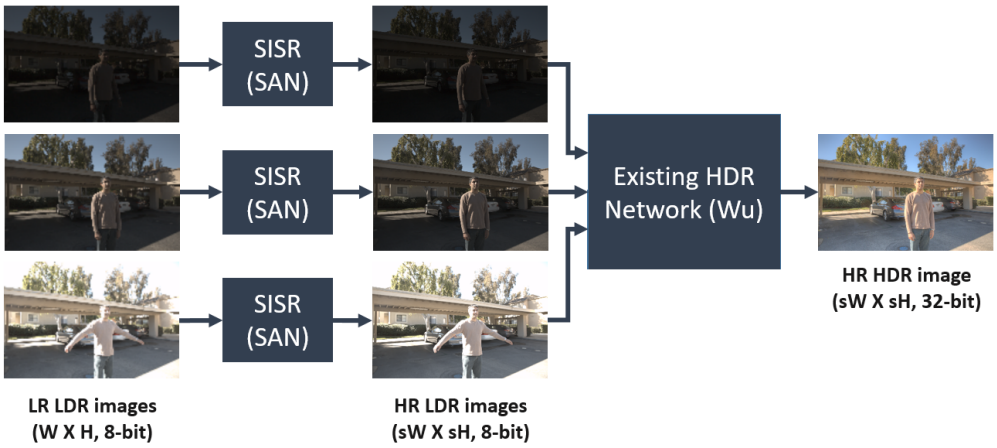
The existing HDR imaging network can also be used for the joint HDR-SR. Among the various HDR methods, Wu [78] is selected and trained for joint HDR-SR task. However, the network in [78] doesn't generate higher-resolution output. Thus, the single image super-resolution (SISR) network in the SAN in [8] should be added before or after the HDR imaging network as shown in Fig. 5.2. The network is trained to generate the final HR and HDR image with three LR and LDR images, likewise in Sec. 5.3.

5.2 Experimental Results

The joint HDR-SR results of proposed network and variations of HDR and VSR networks are compared. In Fig. 5.4 and Fig. 5.5, the reconstructed HR and HDR images of various methods are compared, which are tonemapped with Photomatix to be visu-



(a) Wu+SAN



(b) SAN+Wu

Figure 5.3: The joint HDR-SR scheme using the existing HDR network. The SISR network is added (a) after and (b) before the HDR network, LR and LDR images are fused into an HR and HDR image.

alized. While other methods cause artifacts in the face in Fig. 5.4, and the edge of the tree in Fig. 5.5, the proposed HDR imaging network can reconstruct the HDR images without artifact.

In Table. 5.1, PSNR, SSIM and HDR-VDP-2 between the HDR-SR results and the corresponding ground truth images are computed to evaluate the methods quantitatively. It can be seen that the proposed network provides better performance than the others.

Table 5.1: Comparisons of joint HDR-SR results of the proposed network and variations of HDR and VSR networks. The best results are shown in boldface.

	PSNR-T	SSIM-T	PSNR-L	SSIM-L	HDR-VDP-2
Wu+SAN	33.06	0.8976	32.56	0.8917	54.27
SAN+Wu	33.58	0.9009	33.10	0.8945	54.55
VESPCN	33.93	0.9032	33.21	0.8945	54.68
RBPN	35.55	0.9101	35.27	0.9054	55.08
Proposed	**35.79**	**0.9112**	**35.33**	**0.9098**	**55.52**

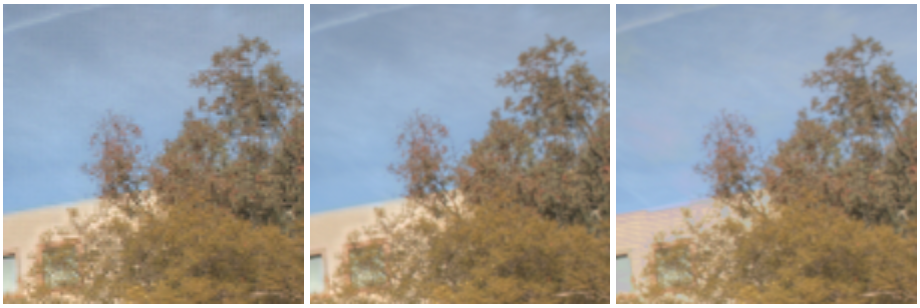
(a) Input(low) (b) Input(mid) (c) Input(high)

(d) Wu [78]+SAN [8] (e) SAN [8]+Wu [78] (f) VESPCN [6]

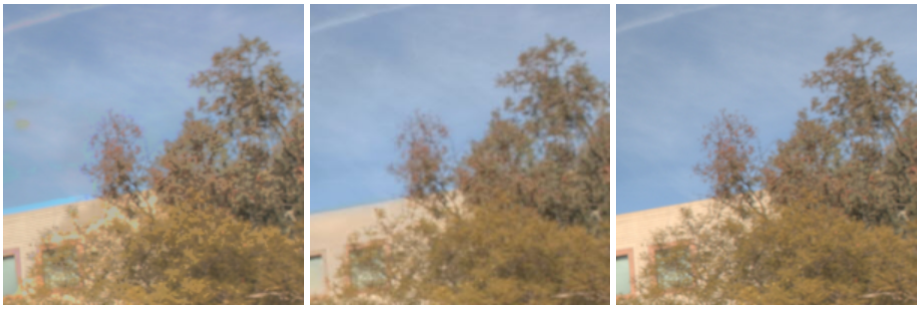
(g) RBPN [21] (h) Proposed (i) GroundTruth

Figure 5.4: Comparsions of the proposed and other methods using existing VSR and HDRI netowork. (a) (b) (c) Input images with different exposures, (d) Wu [78]+SAN [8], (e) SAN [8]+Wu [78], (f) VESPCN [6], (g) RBPN [21], (h) Proposed method and (i) Ground Truth.

(a) Input(low) (b) Input(mid) (c) Input(high)



(d) Wu [78]+SAN [8] (e) SAN [8]+Wu [78] (f) VESPCN [6]



(g) RBPN [21] (h) Proposed (i) GroundTruth

Figure 5.5: Comparsions of the proposed and other methods using existing VSR and HDRI netowork. (a) (b) (c) Input images with different exposures, (d) Wu [78]+SAN [8], (e) SAN [8]+Wu [78], (f) VESPCN [6], (g) RBPN [21], (h) Proposed method and (i) Ground Truth.

Chapter 6

Conclusion

This dissertation have proposed an end-to-end network for the HDR imaging from a set of multi-exposure LDR images with camera and object motions. The conventional deep-learning approaches for the HDR imaging adopted optical flow, homography transform, or attention mechanism to alleviate the problems caused by motions. Unlike the previous methods, a deep network is used to generate aligned images and then the aligned LDR images are merged by another network. The proposed alignment network attempts to synthesize an image that has the exposure of under- or over-exposed image and the structure of standard exposure image, and thus a set of well-aligned LDR images are generated as a result. Then a simple merging network can synthesize a quality HDR image without ghosting or blurring artifacts. The overall network, *i.e.,* the cascade of the alignment and merging networks, is trained end-to-end so that joint learning of alignment-merging is attempted. Extensive experiments show that the proposed network shows better quantitative and qualitative results than state-of-the-art methods. In addition, it is shown that the alignment network effectively works for generating the aligned images with significant illumination differences, and it can be used as a pre-processing step for other HDR imaging methods.

This dissertation also propose an end-to-end network for the VSR from a set of adjacent LR video frames with motions. The conventional deep-learning approaches

for VSR adopted optical flow or STN modules to alleviate the problems caused by motions. Unlike the previous methods, a deep network to generate aligned frames are used and then the aligned LR frames are merged by another network. The proposed alignment network attempts to synthesize an image that has the detail of neighbor frame and the structure of the center frame, and thus a set of well-aligned LR frames are generated as a result. Then a simple reconstruction network can synthesize a quality HR frame without artifacts. The overall network, *i.e.,* the cascade of the alignment and reconstruction networks, is trained end-to-end so that joint learning of alignment-reconstruction is attempted. Extensive experiments show that the proposed network shows better quantitative and qualitative results than state-of-the-art methods.

Furthermore, the proposed VSR network is applied to the joint HDR-SR task. The proposed alignment network attempts to synthesize an image that has the exposure and the detail of under- or over-exposed image and the structure of standard exposure image, and thus a set of well-aligned LDR and LR images are generated as a resu. Then a simple reconslttruction network can synthesize a quality HDR and HR image without artifacts. The overall network, *i.e.,* the cascade of the alignment and reconstruction networks, is trained end-to-end so that joint learning of alignment-reconstruction is attempted. Extensive experiments show that the proposed network shows better quantitative and qualitative results than variations from existing HDR and VSR networks.

Bibliography

[1] Jaehyun An, Seong Jong Ha, and Nam Ik Cho. Reduction of ghost effect in exposure fusion by detecting the ghost pixels in saturated and non-saturated regions. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1101–1104. IEEE, 2012.

[2] Jaehyun An, Seong Jong Ha, and Nam Ik Cho. Probabilistic motion pixel detection for the reduction of ghost artifacts in high dynamic range images from multiple exposures. *EURASIP Journal on Image and Video Processing*, 2014(1):42, 2014.

[3] Jaehyun An, Sang Heon Lee, Jung Gap Kuk, and Nam Ik Cho. A multi-exposure image fusion algorithm without ghost effect. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1565–1568. IEEE, 2011.

[4] Francesco Banterle, Kurt Debattista, Alessandro Artusi, Sumanta Pattanaik, Karol Myszkowski, Patrick Ledda, and Alan Chalmers. High dynamic range imaging and low dynamic range expansion for generating hdr content. In *Computer graphics forum*, volume 28, pages 2343–2367. Wiley Online Library, 2009.

[5] Luca Bogoni. Extending dynamic range of monochrome and color images through fusion. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, pages 7–12. IEEE, 2000.

[6] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4778–4787, 2017.

[7] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018.

[8] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11065–11074, 2019.

[9] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 369–378. ACM Press/Addison-Wesley Publishing Co., 1997.

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[11] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. In *ACM transactions on graphics (TOG)*, volume 21, pages 257–266. ACM, 2002.

[12] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM Transactions on Graphics (TOG)*, 36(6):178, 2017.

[13] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph*, 36(6), 2017.

[14] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):1–11, 2011.

[15] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002.

[16] Xueyang Fu, Delu Zeng, Yue Huang, Yinghao Liao, Xinghao Ding, and John Paisley. A fusion-based enhancing method for weakly illuminated images. *Signal Processing*, 129:82–96, 2016.

[17] Brian Funt and Lilong Shi. The rehabilitation of maxrgb. In *Color and imaging conference*, volume 2010, pages 256–259. Society for Imaging Science and Technology, 2010.

[18] Miguel Granados, Boris Ajdin, Michael Wand, Christian Theobalt, Hans-Peter Seidel, and Hendrik PA Lensch. Optimal hdr reconstruction with linear digital cameras. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 215–222. IEEE, 2010.

[19] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017.

[20] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018.

[21] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2019.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] Yong Seok Heo, Kyoung Mu Lee, Sang Uk Lee, Youngsu Moon, and Joonhyuk Cha. Ghost-free high dynamic range imaging. In *Asian Conference on Computer Vision*, pages 486–500. Springer, 2010.

[24] Jun Hu, Orazio Gallo, and Kari Pulli. Exposure stacks of live scenes with hand-held cameras. In *European Conference on Computer Vision*, pages 499–512. Springer, 2012.

[25] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1163–1170, 2013.

[26] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015.

[27] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Advances in Neural Information Processing Systems*, pages 235–243, 2015.

[28] Yongqing Huo, Fan Yang, Le Dong, and Vincent Brost. Physiological inverse tone mapping based on retina response. *The Visual Computer*, 30(5):507–517, 2014.

[29] Katrien Jacobs, Celine Loscos, and Greg Ward. Automatic high-dynamic range image generation for dynamic scenes. *IEEE Computer Graphics and Applications*, 28(2):84–93, 2008.

[30] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3224–3232, 2018.

[31] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. Properties and performance of a center/surround retinex. *IEEE transactions on image processing*,

6(3):451–462, 1997.

[32] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[33] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017.

[34] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph*, 36(4):144, 2017.

[35] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 319–325. ACM, 2003.

[36] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, 2016.

[37] Kanita Karaduzovic-Hadziabdic, Telalovic J Hasic, and Rafal Konrad Mantiuk. Multi-exposure image stacks for testing hdr deghosting methods. 2017.

[38] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[39] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016.

[40] Ron Kimmel, Michael Elad, Doron Shaked, Renato Keshet, and Irwin Sobel. A variational framework for retinex. *International Journal of computer vision*, 52(1):7–23, 2003.

[41] Rafael P Kovaleski and Manuel M Oliveira. High-quality reverse tone mapping for a wide range of exposures. In *Graphics, Patterns and Images (SIBGRAPI), 2014 27th SIBGRAPI Conference on*, pages 49–56. IEEE, 2014.

[42] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

[43] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[44] Chul Lee, Yuelong Li, and Vishal Monga. Ghost-free high dynamic range imaging via rank minimization. *IEEE Signal Processing Letters*, 21(9):1045–1049, 2014.

[45] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 596–611, 2018.

[46] Sang-hoon Lee, Jae Sung Park, and Nam Ik Cho. A multi-exposure image fusion based on the adaptive weights reflecting the relative pixel intensity and global gradient. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1737–1741. IEEE, 2018.

[47] Dingyi Li and Zengfu Wang. Video superresolution via motion compensation and deep residual learning. *IEEE Transactions on Computational Imaging*, 3(4):749–762, 2017.

[48] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video super-resolution via deep draft-ensemble learning. In *Proceedings of the IEEE International Con-*

ference on Computer Vision, pages 531–539, 2015.

[49] Ce Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009.

[50] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2507–2515, 2017.

[51] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[52] Kede Ma and Zhou Wang. Multi-exposure image fusion: A patch-wise approach. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1717–1721. IEEE, 2015.

[53] Rafał Mantiuk, Scott Daly, and Louis Kerofsky. Display adaptive tone mapping. In *ACM Transactions on Graphics (TOG)*, volume 27, page 68. ACM, 2008.

[54] Rafat Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. In *ACM Transactions on graphics (TOG)*, volume 30, page 40. ACM, 2011.

[55] Belen Masia, Sandra Agustin, Roland W Fleming, Olga Sorkine, and Diego Gutierrez. Evaluation of reverse tone mapping through varying exposure conditions. *ACM transactions on graphics (TOG)*, 28(5):160, 2009.

[56] Belen Masia, Ana Serrano, and Diego Gutierrez. Dynamic range expansion based on image statistics. *Multimedia Tools and Applications*, 76(1):631–648, 2017.

[57] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. In *Computer graphics forum*, volume 28, pages 161–171. Wiley Online Library, 2009.

[58] Peyman Milanfar. *Super-resolution imaging*. CRC press, 2017.

[59] Kamal Nasrollahi and Thomas B Moeslund. Super-resolution: a comprehensive survey. *Machine vision and applications*, 25(6):1423–1468, 2014.

[60] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1219–1232, 2015.

[61] Jae Sung Park, Jae Woong Soh, and Nam Ik Cho. High dynamic range and super-resolution imaging from a single image. *IEEE Access*, 2018.

[62] Jae Sung Park, Jae Woong Soh, and Nam Ik Cho. Generation of high dynamic range illumination from a single image for the enhancement of undesirably illuminated images. *Multimedia Tools and Applications*, 78(14):20263–20283, 2019.

[63] Eduardo Pérez-Pellitero, Jordi Salvador, Javier Ruiz-Hidalgo, and Bodo Rosenhahn. Psyco: Manifold span reduction for super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1837–1845, 2016.

[64] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM transactions on graphics (TOG)*, 21(3):267–276, 2002.

[65] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[66] Guilherme JM Rosa. The elements of statistical learning: Data mining, inference, and prediction by hastie, t., tibshirani, r., and friedman, j. *Biometrics*, 66(4):1315–1315, 2010.

[67] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition, pages 6626–6634, 2018.

[68] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.*, 31(6):203–1, 2012.

[69] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[70] Jae Woong Soh, Jae Sung Park, and Nam Ik Cho. Joint high dynamic range imaging and super-resolution from a single image. *IEEE Access*, 7:177427–177437, 2019.

[71] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[72] Jinshan Tang, Eli Peli, and Scott Acton. Image enhancement using a contrast measure in the compressed domain. *IEEE Signal Processing Letters*, 10(10):289–292, 2003.

[73] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4472–4480, 2017.

[74] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1865–1873, 2016.

[75] Anna Tomaszewska and Radoslaw Mantiuk. Image registration for multi-exposure high dynamic range image acquisition. 2007.

[76] Tsun-Hsien Wang, Cheng-Wen Chiu, Wei-Chen Wu, Jen-Wen Wang, Chun-Yi Lin, Ching-Te Chiu, and Jing-Jia Liou. Pseudo-multiple-exposure-based tone fusion with local region adjustment. *IEEE Transactions on Multimedia*, 17(4):470–484, 2015.

[77] Greg Ward. Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures. *Journal of Graphics Tools*, 8(2):17–30, 2003.

[78] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 117–132, 2018.

[79] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. *IEEE Conference on Compute rVision and Pattern Recognition (CVPR)*, 2019.

[80] Qingsen Yan, Jinqiu Sun, Haisen Li, Yu Zhu, and Yanning Zhang. High dynamic range imaging by sparse representation. *Neurocomputing*, 269:160–169, 2017.

[81] Chih-Yuan Yang, Jia-Bin Huang, and Ming-Hsuan Yang. Exploiting self-similarities for single frame super-resolution. In *Asian conference on computer vision*, pages 497–510. Springer, 2010.

[82] Jianchao Yang, Zhe Lin, and Scott Cohen. Fast image super-resolution based on in-place example regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1059–1066, 2013.

[83] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang. Coupled dictionary training for image super-resolution. *IEEE transactions on image processing*, 21(8):3467–3478, 2012.

[84] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal

correlations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3106–3115, 2019.

[85] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017.

[86] Wei Zhang and Wai-Kuen Cham. Gradient-directed multiexposure composition. *IEEE Transactions on Image Processing*, 21(4):2318–2323, 2012.

[87] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018.

[88] Henning Zimmer, Andrés Bruhn, and Joachim Weickert. Freehand hdr imaging of moving scenes with simultaneous resolution enhancement. In *Computer Graphics Forum*, volume 30, pages 405–414. Wiley Online Library, 2011.

초 록

본 학위논문은 배경 및 전경의 움직임이 있는 상황에서 고 명암비 영상법을 위한 딥 러닝 네트워크를 제안한다. 움직임이 있는 상황에서 촬영된 노출이 다른 여러 영상들을 이용하여 고 명암비 영상을 생성하는 것은 매우 어려운 작업이다. 그렇기 때문에, 최근에 제안된 방법들은 이미지들을 합성하기 전에 패치 매칭, 옵티컬 플로우, 호모그래피 변환 등을 이용하여 그 이미지들을 먼저 정렬한다. 실제로 노출 정도가 다른 여러 이미지들을 정렬하는 것은 아주 어려운 작업이기 때문에, 이 논문에서는 여러 이미지들로부터 얻은 정보를 섞어서 정렬된 이미지를 합성하는 네트워크를 제안한다. 특히, 제안하는 네트워크는 더 밝게 혹은 어둡게 촬영된 이미지들을 중간 밝기로 촬영된 이미지를 기준으로 정렬한다. 주요한 아이디어는 정렬된 이미지를 합성할 때 특징 도메인에서 합성하는 것이며, 이는 픽셀 도메인에서 합성하거나 기하학적 변환을 이용할 때 보다 더 좋은 정렬 결과를 갖는다. 특히, 제안하는 정렬 네트워크는 두 갈래의 인코더와 컨볼루션 레이어들 그리고 디코더로 이루어져 있다. 인코더들은 두 입력 이미지로부터 특징을 추출하고, 컨볼루션 레이어들이 이 특징들을 섞는다. 마지막으로 디코더에서 정렬된 이미지를 생성한다. 제안하는 네트워크는 고 명암비 영상법에서 사용될 수 있도록 노출 정도가 크게 차이나는 영상에서도 잘 작동한다. 게다가, 간단한 병합 네트워크를 추가하고 전체 네트워크들을 한 번에 학습함으로서, 최근에 제안된 방법들 보다 더 좋은 성능을 갖는다.

또한, 본 학위논문은 동영상 내 프레임들을 이용하는 비디오 고 해상화 방법을 위한 딥 러닝 네트워크를 제안한다. 동영상 내 인접한 프레임들 사이에는 움직임이 존재하기 때문에, 이들을 이용하여 고 해상도의 프레임을 합성하는 것은 아주 어려

운 작업이다. 따라서, 최근에 제안된 방법들은 이 인접한 프레임들을 정렬하기 위해 옵티컬 플로우를 계산하거나 STN을 추가한다. 움직임이 존재하는 프레임들을 정렬하는 것은 어려운 과정이기 때문에, 이 논문에서는 인접한 프레임들로부터 얻은 정보를 섞어서 정렬된 프레임을 합성하는 네트워크를 제안한다. 특히, 제안하는 네트워크는 이웃한 프레임들을 목표 프레임을 기준으로 정렬한다. 마찬가지로 주요 아이디어는 정렬된 프레임을 합성할 때 특징 도메인에서 합성하는 것이다. 이는 픽셀 도메인에서 합성하거나 기하학적 변환을 이용할 때 보다 더 좋은 정렬 결과를 갖는다. 특히, 제안하는 정렬 네트워크는 두 갈래의 인코더와 컨볼루션 레이어들 그리고 디코더로 이루어져 있다. 인코더들은 두 입력 프레임으로부터 특징을 추출하고, 컨볼루션 레이어들이 이 특징들을 섞는다. 마지막으로 디코더에서 정렬된 프레임을 생성한다. 제안하는 네트워크는 인접한 프레임들을 잘 정렬하며, 비디오 고해상화에 효과적으로 사용될 수 있다. 게다가 병합 네트워크를 추가하고 전체 네트워크들을 한 번에 학습함으로서, 최근에 제안된 여러 방법들 보다 더 좋은 성능을 갖는다.

고 명암비 영상법과 비디오 고 해상화에 더하여, 본 학위논문은 명암비와 해상도를 한 번에 향상시키는 딥 네트워크를 제안한다. 앞에서 제안된 두 네트워크들은 각각 명암비와 해상도를 향상시킨다. 하지만, 그들은 하나의 네트워크를 통해 한 번에 향상될 수 있다. 이 논문에서는 비디오 고해상화를 위해 제안한 네트워크와 같은 구조의 네트워크를 이용하며, 더 높은 명암비와 해상도를 갖는 최종 결과를 생성해낼 수 있다. 이 방법은 기존의 고 명암비 영상법과 비디오 고해상화를 위한 네트워크들을 조합하는 것 보다 정성적으로 그리고 정량적으로 더 좋은 결과를 만들어 낸다.

주요어: 영상 정합, 영상 정렬, 컨볼루션 신경망, 고 명암비 영상법, 비디오 고 해상화
학번: 2013-20846